Alma Mater Studiorum · Università di Bologna

SCUOLA DI SCIENZE

Corso di Laurea in Informatica

Linked Data e bibliometriche: un indice di multidisciplinarietà nel Semantic Publishing

Relatore:
Chiar.mo Prof.
Fabio Vitali

Presentata da:
Andrea Bagnacani

Correlatore:
Chiar.mo Prof.
Michele Costa

Sessione III
Anno Accademico 2012/2013

Indice

Introduzione

Nel presente lavoro si introduce un nuovo indice per la valutazione dei prodotti della ricerca: l'*indice di multidisciplinarietà*.

Questa nuova metrica può essere un interessante parametro di valutazione: il panorama degli studi multidisciplinari è vasto ed eterogeneo, ed all'interno di questo sono richieste necessarie competenze trasversali.

Le attuali metriche adottate nella valutazione di un accademico, di un journal, o di una conferenza non tengono conto di queste situazioni intermedie, e limitano la loro valutazione dell'impatto al semplice conteggio delle citazioni ricevute.
Il risultato di tale valutazione consiste in un valore dell'impatto della ricerca senza una connotazione della direzione e della rilevanza di questa nel contesto delle altre discipline.

L'indice di multidisciplinarietà proposto si integrerebbe allora all'interno dell'attuale panorama delle metriche di valutazione della ricerca, offrendo -accanto ad una quantificazione dell'impatto- una quantificazione della varietà dei contesti disciplinari nei quali si inserisce.

La disponibilità di metriche che aiutino a valutare l'impatto, così come la qualità dei contributi presentati dagli eventi e dall'editoria della ricerca, si stanno rivelando uno strumento indispensabile all'individuazione di nuovi trend della ricerca.

La possibilità di osservare questo complesso panorama attraverso un modello semplificato del suo evolvere nel tempo influenza decisioni cruciali, come l'allocazione di investimenti o l'assunzione di nuovo personale.

A questo scopo, nel tempo sono stati introdotti diversi indici di valutazione della ricerca, ognuno con il suo raggio d'azione: dalla prima proposta del *Science Citation Index (SCI)* avanzata da Garfield [GS63; Gar72], e focalizzata sulla valutazione dell'impatto dei journal della ricerca e ristretta ad un preciso elonco di journal, alla proposta dell'*h-index* di Hirsh [Hir05], più generalizzata, ed applicabile sia ai prodotti della ricerca sia ai suoi autori.
Il minimo comune denominatore di queste metriche è il conteggio del numero di *citazioni ricevute* da altri prodotti di ricerca.

Collezionare citazioni, o avere accesso a database che forniscano questa informazione, è da allora indispensabile ai fini del raggiungimento di una valutazione della ricerca.

Parallelamente allo sviluppo di indici di impatto, sono stati molti anche gli interrogativi posti in merito alla classificazione di queste citazioni. In particolare, il tema della contestualizzazione della ricerca all'interno di una disciplina (o insieme di discipline) è stato affrontato con l'adozione di una tassonomia delle discipline della ricerca: collocando cioè ogni pubblicazione all'interno di una determinata categoria.

Sono quindi nati portali della ricerca specializzati in questo tipo di operazione, una soluzione che però non riesce a mettere d'accordo tutto il mondo della ricerca, che percepisce le attuali tassonomie come poco precise: molto generiche, o addirittura errate.

Il problema è percepito innanzitutto da quelle comunità di ricerca il cui nome esatto o di uso comune non è documentato dalle tassonomie più note, e soprattutto all'interno di quegli ambienti di ricerca a cavallo tra diverse aree di indagine, dove l'interdisciplinarità è diventata un requisito per la *buona*

ricerca: Biotecnologie e Nanotecnologie, ma anche Bioinformatica, Design, Sociologia, Antropologia, la cui ricerca viene alle volte catalogata con il termine impreciso "Studi Interdisciplinari".

L'adozione di una tassonomia rigida, di scarsa duttilità all'inclusione di aree di ricerca recenti o di nicchia, e fragile a causa dell'inclusione di termini opachi, è quindi da abbandonarsi in favore di soluzioni che siano in grado di comprendere l'affinità dei prodotti della ricerca attraverso le loro interconnessioni.

Una prima risposta a tale esigenza, ed a una caratterizzazione più puntuale del contesto citazionale di una data pubblicazione, è stata introdotta da Leydesdorff [Ley09a], il quale ha sviluppato le sue osservazioni secondo un approccio *bottom-up*, organizzando le discipline non secondo una mappa concettuale del sapere noto, ma attraverso alcune proprietà della rete citazionale in esame, ed assegnando dei valori di somiglianza ed affinità agli insiemi individuati.
Sebbene Leydesdorff riconducesse questi insiemi alle discipline della tassonomia *ISI*[1], l'idea di base dello sfruttamento delle proprietà della rete è un'intuizione che trova fondamento ed applicazioni in un'altra ampia area di studi: la *Teoria delle Reti* [EK10].

La valutazione della multidisciplinarietà dei prodotti della ricerca avanzata nel presente lavoro abbandona quindi l'adozione di una tassonomia delle discipline, e procede secondo una direzione di indagine bottom-up come quella proposta da Leydesdorff, ma condotta sfruttando appieno le proprietà delle reti *Small World* delineate da Watts, Strogatz, e Kleinberg [WS98; Kle00; EK10].

Secondo questo nuovo paradigma di studio delle reti citazionali, ogni pub-

[1] http://thomsonreuters.com/

blicazione è interpretata con un nodo, mentre ogni citazione con un arco direzionato tra pubblicazione citante e pubblicazione citata.

In una rete Small World, il numero previsto di salti tra una coppia di nodi scelti a caso è piccolo secondo una certa definizione di scala. In particolare, è Small World se la lunghezza media del cammino minimo tra coppie di nodi D rispetto al numero totale di nodi N nella rete rispetta la funzione:

$$D = K \log N$$

Secondo questo fattore di crescita è presente un'elevata interconnessione tra nodi affini (le comunità omogenee), e bassa interconnessione tra comunità eterogenee.
In questo modo:

- i documenti monodisciplinari di un dato contesto disciplinare condividono in media più link (citazioni) da e per i documenti a loro affini;

- i documenti multidisciplinari condividono un numero inferiore di link da e per i documenti a loro affini, svolgendo una funzione di "ponte" tra le varie comunità (le discipline) della rete in esame.

Le comunità disciplinari in una rete Small World sono quindi una naturale proprietà di questa tipologia di grafo.

Nel presente lavoro, la valutazione della multidisciplinarietà di un prodotto della ricerca adotta quindi questo paradigma, rimodellando il grafo citazionale in una rete Small World, ed analizzandolo per mezzo di un algoritmo di clustering.

Allo scopo di testare questo approccio bottom-up, e fornire una valutazione della multidisciplinarietà dei prodotti della ricerca, si è resa necessaria la costruzione di una rete citazionale.
Questo compito è svolto da un nuovo strumento software: *CoData*.

CoData, "Context/Community Data" è un software che fornisce una descrizione dei prodotti della ricerca, delle loro relazioni, e del loro contesto. Inoltre offre funzionalità di valutazione della ricerca attraverso il calcolo di metriche di impatto, e di multidisciplinarietà.

Nel presente lavoro, CoData costruisce la rete citazionale a partire dal contenuto di due dataset: *DBLP*[2], che contiene record di pubblicazioni relative a journal, conferenze, e proceeding relativi al mondo delle Scienze dell'Informazione, e *DBLP+C*[3], i cui record sono analoghi a quelli di DBLP, ma la cui rete citazionale è molto più estesa ed aggiornata.

Data la diversità con la quale i due dataset formalizzano i propri record, CoData provvede ad una prima fase di *data retrieval* e *data integration*, con la quale:

- esprime ogni record in un modello semantico, utilizzando una famiglia di ontologie del *Semantic Publishing*: SPAR [PS12; PSV12], SWC [DER], e SWRC [Sur+05];
- sfrutta il modello di pubblicazione *Linked Data* creando una *Knowledge Base* dei record di entrambi i dataset.

I dati così riorganizzati sono memorizzati in un triple-store *Virtuoso Open Source (VOS)*[4] accessibile mediante uno SPARQL end-point.

L'individuazione delle comunità (e delle discipline) nella rete citazionale comincia con la presenza del grafo delle pubblicazioni della ricerca. CoData crea da questi il *Connected Component*: una rete in cui ogni coppia di nodi è collegata da un cammino di archi citazionali, la cui struttura riflette quella del grafo Small World descritto in precedenza.

[2]http://dblp.uni-trier.de/xml/
[3]http://arnetminer.org/citation
[4]http://virtuoso.openlinksw.com/dataspace/doc/dav/wiki/Main/

Attraverso questo nuovo modello dei dati, l'individuazione delle comunità (le discipline) della ricerca è possibile mediante l'uso di un algoritmo di clustering.

Nel presente lavoro questa operazione è affidata a *Chinese Whispers*: un algoritmo di hard clustering che non ha bisogno di parametri di configurazione, e che ottiene buone performance su grafi Small World di grandi dimensioni [Bie06].

CoData svolge quindi tre funzioni:

- collezione e pubblicazione dei dati relativi alla ricerca raccolta in due dataset;
- costruzione e analisi delle proprietà della rete citazionale risultante;
- valutazione della multidisciplinarietà della ricerca.

Quest'ultima funzionalità è realizzata formalizzando la definizione di *multidisciplinarietà*.

Si definisce *multidisciplinarietà di un prodotto della ricerca* una caratteristica di disciplinarità *attribuita* dalle comunità disciplinari della rete citazionale contestuale all'opera considerata. Un'opera è quindi multidisciplinare quando le citazioni che riceve le attribuiscono rilevanza nel contesto di più ambiti disciplinari.

Secondo questa definizione, la multidisciplinarietà dipende dalle caratteristiche di *rilevanza* e *cardinalità disciplinare* del contributo, mentre è invece indipendente dall'impatto che questa ottiene nella disciplina citante.

Per poter arrivare ad una quantificazione della multidisciplinarietà di un opera, CoData analizza la distribuzione delle citazioni tra le discipline della rete, ed assegna ad ogni coppia una *soglia di rilevanza* citazionale. Questa definisce un numero di citazioni minimo che il contributo deve aver ricevuto

per poter essere considerato disciplinato presso la disciplina citante.

I tempi medi di realizzazione di ogni operazione variano a seconda delle tre fasi di costruzione sopra elencate. In particolare, durante la raccolta e arricchimento dei dati utile alla costruzione del triple-store, CoData è efficiente nell'operazione di trasformazione dei record relativi agli autori (qualche centesimo di secondo), mentre meno efficiente relativamente all'integrazione dei record tra i dataset DBLP e DBLP+C (circa 6 secondi).
Un aumento considerevole dei tempi di esecuzione che si spiega con l'elevato numero di operazioni di interrogazione e pattern-matching sulle stringhe del triple-store, necessarie alla verifica di una buona qualità di integrazione tra record rappresentati da identificativi differenti.

Qualitativamente, l'integrazione dei dettagli della pubblicazione con il modello semantico degli autori, la publication venue, e la rete citazionale ricavata dal dataset DBLP+C, portano all'ottenimento di un modello semantico molto più chiaro e dettagliato del semplice record taggato con stringhe arbitrarie.
Ai vantaggi di una più puntuale rappresentazione dei dettagli di un prodotto della ricerca, si aggiunge quella della disponibilità degli stessi nel Linked Data: i record DBLP e DBLP+C sono infatti riusabili, ed identificabili univocamente nel *World Wide Web*.

CoData è efficiente anche durante la clusterizzazione del modello Small World, il quale impiega -in media- circa 5 minuti per l'assegnazione di 56'157 discipline a 380'527 pubblicazioni, mentre -in media- circa 6 secondi per la costruzione della matrice delle distribuzioni citazionali in un insieme di 712 discipline (circa l'1% del totale delle pubblicazioni in cluster disciplinari con almeno cinquanta documenti).

Questi risultati offrono una visione d'insieme relativa ai dati disponibili nei dataset DBLP e DBLP+C, la quale indica che la maggioranza dei cluster

contiene un numero di pubblicazioni molto piccolo. Questo dà adito a due osservazioni:

- le pubblicazioni all'interno degli insiemi disciplinari non condividono molte citazioni con il resto della rete citazionale. Ad esempio, i nodi di questi cluster potrebbero rappresentare pubblicazioni che citano, ma che non vengono mai citate;
- i record dei dataset utilizzati mancano di alcune voci di bibliografia.

Il presente lavoro è quindi solo l'inizio di un progetto di indagine più ampio, e che tocca nuove prove, e nuovi miglioramenti nelle tecnolgie proposte, e negli algoritmi utilizzati.
In particolare, tra gli sviluppi futuri di CoData si suggeriscono:

- l'ottimizzazione del modello semantico dei dati, ora ricco di dettagli relativi ai record dei dataset DBLP e DBLP+C, ma ridondante nella descrizione della rete citazionale;
- l'integrazione di nuovi dataset, come *Scopus*[5], e *ACM*[6];
- la costruzione di un authority file per il mondo della ricerca, in modo da ottimizzare il processo di integrazione dei dati, evitando operazioni di pattern-matching, e sfruttando appieno le potenzialità del Linked Data;
- il test di altri algoritmi di clustering, ad esempio K-Means, o algoritmi di clustering gerarchici, che ricalchino la struttura di una tassonomia delle discipline, gerarchizzando le relazioni rappresentate nel grafo citazionale.

[5]http://www.elsevier.com/
[6]http://acm.rkbexplorer.com/sparql/

Capitolo 1

Discipline scientifiche e metriche citazionali

La disponibilità di metriche che aiutino a valutare la produttività e l'impatto di un ricercatore, così come la bontà e la qualità dei contributi presentati da journal, conferenze, e workshop, si stanno rivelando uno strumento indispensabile all'individuazione di nuovi trend nel mondo della ricerca: la possibilità di osservare questo complesso panorama attraverso un modello semplificato del suo evolvere nel tempo influenza decisioni cruciali, come la direzione di indagine di una ricerca, l'allocazione di investimenti, o l'assunzione di nuovo personale.

In questo capitolo si fornisce una panoramica sullo stato dell'arte delle metriche per la valutazione dei prodotti della ricerca.

La necessità di quantificare la qualità e l'evolversi della ricerca hanno dato vita a diversi metodi di valutazione. In particolare, Pritchard [Pri69] definisce con il termine *Bibliometrics* (*bibliometriche*), l'insieme delle metodologie della matematica e della statistica volte all'analisi ed alla quantificazione della letteratura della ricerca, e ancora, con il termine di *Scientometrics* si intende l'insieme delle tecniche per l'analisi e la quantificazione delle attività

scientifiche (includendo nell'elenco anche l'organizzazione di libri e journal), mentre con *Informetrics* si definisce l'insieme delle tecniche per entrambe le sfere: quella scientifica, e quella della letteratura in generale.

Dopo Pritchard, lo studio di nuove bibliometriche ha preso sempre più piede.

La proposta dell'indice di impatto *Science Citation Index (SCI)* avanzata da Garfield [GS63; Gar72] ad esempio, prevede la quantificazione della bontà della ricerca presentata dai maggiori journal dell'epoca; una proposta a lungo dibattuta [Sol70; CN73; Gil77; Nar76; Ben01], soprattutto a fronte dell'osservazione che differenti discipline tendono ad avere differenti modi di citare e citarsi.
Di fronte a queste polemiche, SCI si è trasformato in un ambiente di indici che tiene conto non solo delle citazioni ricevute da un journal, ma anche dalla sua disciplina; un termine di valutazione che necessita l'introduzione di categorie disciplinari per ogni journal.

A questo scopo, *Institute of Science Index (ISI)*[1] introduce un metodo di valutazione della ricerca che prevede la creazione di una *tassonomia delle discipline* [LB09]: una classificazione gerarchica degli ambiti di ricerca, ottenuta collezionando i "termini disciplinari" presenti nel titolo dei journal, e riorganizzandoli gerarchicamente.

I metodi sviluppati fino a quel punto non sono però soddisfacenti: Leydesdorff [LB09] evidenzia che questi non siano equi, in quanto non è possibile applicare un indice di impatto a due journal distinti ma appartenenti alla stessa categoria tassonomica senza prendere in considerazione anche la loro età.
A fronte di questo vincolo restrittivo, Leydesdorff suggerisce di relazionare il numero delle citazioni ricevute alla *interdisciplinarità* del journal in esame.

[1] http://thomsonreuters.com/

È necessario allora porre il numero di citazioni all'interno di un contesto più ampio ed articolato, misurabile in modo automatico, e valutabile indipendentemente dall'età della ricerca.
Si vuole osservare la citazione in un ambiente che sappia cogliere aspetti come la *connettività*, l'*interazione*, e l'*affinità* tra *comunità disciplinari*.

Un metodo per l'identificazione di comunità disciplinari adatto allo studio di tali caratteristiche, si ritrova nella *Teoria delle Reti*, per la quale questi aspetti sono classificati come una proprietà intrinseca della rete citazionale dei prodotti della ricerca [WS98; EK10].
Da questi studi emerge infatti che un particolare contesto è effettivamente in grado di influenzare fortemente il posizionamento di nuovi *link* (citazioni), non solo, ma di determinarne anche direzione, verso, ed intensità.

In una rete citazionale dunque, il contesto di una citazione è in grado non solo di spiegare la presenza e la direzione della citazione stessa, ma anche l'emergere di nuove comunità disciplinari, e l'evolvere di nuovi trend.

1.1 Una categorizzazione delle discipline

Di seguito si illustrano i metodi adottati nella categorizzazione delle discipline: dalle tassonomie della ricerca, all'analisi delle proprietà della rete citazionale.

1.1.1 Una tassonomia delle discipline

Le tassonomie delle discipline sono molteplici, e ricoprono diverse aree dell'indagine scientifica. Alcune di queste di ampia condivisione nel mondo della ricerca, altre meno.
In questa sezione si introduce il panorama delle tassonomie delle discipline, le categorie della ricerca, ed alcuni metodi alternativi alla definizione di tali gerarchie.

La tassonomia delle discipline della ricerca introdotta da ISI (oggi *Thomson Reuters Web of Knowledge, o WoK*[2]) è costruita in modo automatico, collezionando alcuni termini del titolo di journal e proceedings, e disponendoli gerarchicamente secondo alcuni criteri di ricorrenza [LB09].
Sebbene si sia dibattuto a lungo sulla sua completezza ed espressività [Ben01], è ancora uno strumento molto sfruttato, soprattutto per il contesto nel quale va ad inserirsi: WoK.

Ad oggi WoK è molto più di un semplice elenco di discipline a disposizione dei ricercatori: è un vero e proprio portale col quale -previa sottoscrizione di un abbonamento- è possibile accedere ad un elevato numero di servizi, come il download dell'elenco delle discipline, e milioni di documenti, tra pubblicazioni, abstract, statistiche, e bibliometriche relative a più di 16.000 journal reference.
Attraverso questa preziosa base di conoscenza è stato possibile condurre una serie di esperimenti per la visualizzazione grafica di una mappatura delle discipline [LCR12] e delle loro interazioni [Zha+10] che, per quanto non esaustiva o imprecisa (a causa della presenza di alcune categorie troppo opache [Zit05]), offre un'idea chiara dell'estensione e dell'interazione della conoscenza dei nostri giorni.

La classificazione ISI non è l'unica tassonomia delle discipline della ricerca: tra le sue dirette concorrenti vi sono ad esempio le classificazioni *ACM Computing Classification System*[3] del portale di *Association for Computing Machinery (ACM) Digital Library*[4] (aggiornata di recente [Cou97]), e la classificazione *Elsevier*[5] (di recente introduzione).

La costruzione di una tassonomia non è tuttavia l'unica strada percorsa all'individuazione di un insieme di discipline, in quanto i costi di manteni-

[2]http://wokinfo.com/
[3]http://www.acm.org/about/class/
[4]http://dl.acm.org/
[5]http://www.elsevier.com/

mento per tale operazione linguistica sono alti e virtualmente costanti nel tempo. In secondo luogo il suo continuo aggiornamento ai trend della ricerca deve riflettere i cambiamenti di un mondo che progredisce ad una velocità molto più sostenuta di quella necessaria per un riadattamento che rifletta la presenza di nuove correnti, e nuove comunità disciplinari. Inoltre manca spesso un accordo tra gli stessi ricercatori circa quali siano i confini di una disciplina, o i confini della propria ricerca!

Ai costi della definizione di una tassonomia ed alla mancanza d'accordo tra i ricercatori, si aggiunge un altro problema: quello delle categorie opache. Un esempio tra tutti è dato dalla categoria *Studi Interdisciplinari*.

Con "Studi Iterdisciplinari" si identificano tutti quegli studi proposti all'integrazione di conoscenze trasversali, che uniscono discipline (e ricercatori) di diversi ambiti o diversi approcci d'indagine; ambiti che trovano un punto di affinità nello studio di soggetti, la cui analisi risulta più comprensibile attraverso l'integrazione dei background disciplinari [Sci04].

La genericità di questo termine tassonomico, già dibattuta in passato [Zit05], è un vero e proprio ostacolo alla comprensione e valutazione del mondo della ricerca, e produce due effetti negativi sullo studio del contesto della citazione:

- Rafols e Meyer [RM10] osservano che mentre journal di una stessa disciplina presentano caratteristiche e schemi citazionali simili, journal catalogati con l'etichetta "Studi Interdisciplinari" non seguono questo comportamento, evidenziando così la poca affinità tra journal di questa categoria tassonomica

- l'ambiguità del termine rende la tassonomia estremamente fragile, rendendone di fatto rischioso l'impiego nell'analisi di un corpo documentale

Il concetto di "interdisciplinarità", come inteso nel contesto degli Studi Interdisciplinari, è dunque un termine troppo opaco [RM10] e di difficile adozione

per una valutazione equa della ricerca.

Un approccio alternativo alle tassonomie consiste nel riferirsi a parole chiave e *buzz-word* che identifichino nomi di discipline, o particolari termini significativi alla ricerca ed ai suoi trend.

L'insieme di questi termini preferenziali definiscono quello che in gergo prende il nome di *folksonomia* [Wal04]: una "tassonomia del popolo", nella quale l'identificazione di una disciplina o di un trend è strettamente dipendente dalla prevalenza statistica di certi termini piuttosto che altri.
Quest'approccio, più automatico all'ottenimento di un elenco delle discipline, non è vantaggioso solo per la comodità derivata dall'assenza del mantenimento di una tassonomia e del suo vocabolario controllato, ma anche perché permette l'inclusione di termini più recenti.

La disponibilità di folksonomie ha permesso diverse applicazioni nell'ambito del *Machine Learning*, in particolare nello studio dell'individuazione di trend nella ricerca attraverso l'analisi di una rete citazionale, osservando la distribuzione delle citazioni ricevute nel contesto dei termini folksonomici.

A tale proposito, Jun, Park, e Jang [JPJ14] propongono l'individuazione di nuovi trend della ricerca, collezionando non le sole parole chiave di un set di pubblicazioni, ma il loro intero corpo del testo.
I termini da loro raccolti sono riorganizzati in *feature vectors* ovvero righe di una matrice m x n dove m è il numero di documenti raccolti, ed n è lo *spazio* di tutti i termini di ogni pubblicazione.

La matrice ottenuta è una matrice sparsa dei termini di una pubblicazione, i cui valori indicano la frequenza d'apparizione nel loro contesto (quello della pubblicazione d'appartenenza).
Per risolvere il problema della *sparseness* della matrice, ed individuare i termini più trendy al suo interno, Jung, Park, e Jang usano invece un metodo combinato di riduzione matriciale (*PCA* [DH04]) e clustering (*SVC* [Ben+02]).

Il risultato da loro ottenuto, testato con un dataset di prova, si è rivelato molto promettente, e non è da escludere che con il tempo, queste tecniche combinate di Machine Learning e cluster analysis [Sch07] prendano sempre più piede nello studio dei trend della ricerca.

In questa sezione si è illustrato come un approccio linguistico *top-down* (tassonomico/folksonomico) dia un'idea chiara di come le discipline ed i trend della ricerca siano organizzati e relazionati.
Sebbene questi metodi introducano un sapere prezioso per la definizione e visualizzazione di una mappa delle discipline, la tassonomia si rivela costosa da mantenere, di scarsa duttilità all'inclusione di aree di ricerca recenti o di nicchia, e molto fragile a causa dell'inclusione di termini opachi; mentre le folksonomie sono mutevoli, ridondanti, prone a errori, e poco adatte ad una rappresentazione puntuale del contesto disciplinare di una pubblicazione.

1.1.2 Le discipline secondo indici di diversità

Di seguito si introducono alcune tecnologie che permettono un processo di identificazione delle discipline della ricerca. Tale processo è svincolato dall'adozione di particolari tassonomie o termini folksonomici, e basato sullo studio delle proprietà di una rete citazionale.

In uno studio di una rete citazionale interdisciplinare nel campo delle *Nanoscienze* e *Nanotecnologie*, Rafols e Meyer propongono un'indagine in due direzioni [RM10]: nella prima, *top-down*, analizzano la rete citazionale nelle sue tre caratteristiche di *diversità*, *varietà*, e *bilanciamento* [Sti98; Sti07], confrontando successivamente queste valutazioni con i risultati di una seconda indagine *bottom-up*, condotta attraverso interviste dirette ai ricercatori, le cui pubblicazioni costituivano parte della rete citazionale interdisciplinare.

Questo studio evidenzia come differenti journal etichettati con il termine di "Studi Interdisciplinari" presentino tra loro sostanziali differenze in termini di diversità delle discipline trattate, aggiungendo così un'altra prova

all'inadeguatezza di questo termine.
A seguito delle interviste ai ricercatori invece, è emerso come l'interazione tra questi abbia di fatto individuato nuove comunità disciplinari nella rete citazionale (rilevabili dagli indici di *diversità*, *varietà* e *bilanciamento*), ma che queste siano rimaste invisibili alle categorie ISI.
In conclusione, la tassonomia ISI si rivela inadatta a catturare il dinamismo dell'interazione tra discipline in ambiti di ricerca così mutevoli e specializzati.

In un altro studio, Leydesdorff [Ley09a] si basa invece sull'utilizzo delle proprietà delle reti (in particolare le reti sociali) e, prendendo in prestito dalla *Sociologia* il concetto di *centralità* [Fre77], analizza l'interdisciplinarità di un journal in base alla posizione che questi occupa all'interno del cluster disciplinare: l'occupazione di una posizione di centralità o di periferia rispetto al centro ideale del cluster, ne determina il grado di interdisciplinarità.

La sua indagine si conclude con la supposizione che una misura della *multidisciplinarietà* di un journal potrebbe essere dipendente dalla vicinanza che questi ha rispetto ai journal citanti.
Si noti come la sua supposizione riguardi i journal *citanti*, e non i *citati*, evidenziando così che la multidisciplinarietà sia interpretabile come un'informazione da ricercarsi nelle citazioni entranti, e non quelle uscenti. Un'interpretazione che attribuisce al soggetto multidisciplinare una posizione centrale in un background disciplinare variegato, in grado di "tirare a sé" prodotti della ricerca differenti.

Questo tipo di esperimento non si limita ad introdurre al ricercatore una mappa visuale delle discipline della ricerca, ma porta anche all'osservazione che l'individuazione delle discipline possa essere condotta attraverso l'identificazione di caratteristiche proprie della rete citazionale, sfruttando algoritmi in grado di "capire" e "distinguere" queste caratteristiche in modo automatico, senza cioè sfruttare particolari termini di categorizzazione.

In questa sezione si è illustrato come un differente approccio *bottom-up*

possa mettere in risalto particolari proprietà della rete, in grado di descriverne naturalmente la topologia, senza fare affidamento su particolari tassonomie o folksonomie.

1.2 Le metriche citazionali

Di seguito si illustrano alcune tra le più usate metriche per la valutazione della ricerca, assieme ai portali, ed ai dataset presso le quali queste siano reperibili (o deducibili).

1.2.1 La valutazione della ricerca

Ad oggi, la valutazione della ricerca scientifica avviene principalmente secondo due approcci:

- l'approccio *qualitativo*, tra cui ad esempio la pratica del *peer reviewing* (la "valutazione dei pari"), molto importante all'accertamento della bontà e brevettabilità della ricerca;

- l'approccio *quantitativo*, funzionale alla valutazione dell'impatto di una ricerca in termini numerici. Un esempio si ritrova negli indici bibliometrici, studiati esattamente a questo scopo.

Anche se generalmente accettata, l'assunzione che ad un'elevata frequenza citazionale corrisponda un'elevata qualità della rivista, sul piano concettuale, questo ha scarso fondamento [Nar76].
Resta tuttavia pratica abituale riferirsi a metriche quantitative per avere un'idea della bontà della ricerca condotta, ed è a tal proposito che risulta necessario un indicatore (o un insieme di indicatori) in grado di catturare più aspetti sulla conduzione della ricerca, senza limitare il proprio raggio di osservazione al mero conteggio delle citazioni ricevute.

Di seguito si riporta un elenco dei più famosi indici bibliometrici unitamente alle loro evoluzioni ed adattamenti.

Impact Factor (IF) : introdotto da Garfield [GS63; Gar72], l'IF è l'indicatore bibliometrico più conosciuto, e viene calcolato su tutti i journal appartenenti al *Journal Citation Reports (JCR)* come il numero di citazioni ricevute che ogni *citable item*[6] del journal in esame riceve entro i due anni dalla prima pubblicazione nel circuito della ricerca.
In particolare, per calcolare l'IF di un journal pubblicato nel 2014 (ed appartenente all'insieme dei journal di JCR) si considerano

A : il numero di citazioni ricevute che i citable item del journal in esame hanno ricevuto negli anni 2012 e 2013 da journal indicizzati dal JCR nel 2014

B : il numero di citable items pubblicati dal journal in esame negli anni 2012 e 2013

dunque

$$IF_{2014} = \frac{A}{B}$$

IF divenne *standard de facto* nella valutazione della qualità della ricerca, al punto che Garfield stesso mise in guardia dall'adottare IF utilizzandolo come termine di paragone unico [Gar05; Gar06]: in particolare, IF non è calcolabile al di fuori dell'indice dei journal di JCR. Inoltre, come osservato già da vari studi [Sol70; CN73; Gil77; Nar76], differenti discipline tendono ad avere differenti modi di citare e citarsi: journal di discipline mediche possono infatti ottenere valori di IF ben più alti di journal di Scienze dell'Informazione, e tuttavia valere meno in termini qualitativi.
Non solo: IF è applicabile solo al livello dei "citable item" di un journal, restando dunque insensibile alle pubblicazioni fuori rivista, ed all'intero panorama delle conferenze.

Immediacy Index : a fronte delle critiche sollevate in merito all'Impact Factor, ISI pose IF all'interno di un contesto più ampio di indicatori

[6] per "citable item" si intende un articolo, una review, un proceeding, una nota; ma non un editoriale o una comunicazione all'editore

di valutazione, sempre all'interno dei journal appartenenti all'indice di JCR; uno di questi fu appunto l'Immediacy Index: un Impact Factor del breve periodo.
Il suo calcolo è molto simile a quello di IF: per calcolare l'Immediacy Index di un journal pubblicato nel 2014 (ed appartenente all'insieme dei journal di JCR) si considerano:

A : il numero di citazioni ricevute che i citable item del journal in esame hanno ricevuto nel 2014 da journal indicizzati dal JCR nel 2014

B : il numero di citable items pubblicati dal journal in esame nel 2014

dunque

$$ImmediacyIndex_{2014} = \frac{A}{B}$$

Citing Half-Life : la mediana dell'età dei citable item citati dagli articoli pubblicati nel journal in esame durante l'anno di pubblicazione

Rate of Cites Index : basato sull'assioma che tanto più un articolo venga citato, tanto più questi sia rilevante all'interno di un dato ambito di ricerca, il Rate of Cites Index rappresenta non più una valutazione della qualità del journal nell'insieme, ma una valutazione del singolo lavoro.

Il Rate of Cite Index è stato l'unico indice introdotto da ISI con l'obbiettivo di valutare la qualità del singolo.
Il suo limite d'applicazione all'interno della cerchia dei journal registrati presso il JCR, congiuntamente alla sua evidente fragilità (si pensi ad eventuali anniversari o ricorrenze dalla prima pubblicazione dell'articolo considerato: eventi simili gonfierebbero il valore del Rate of Cite Index, spostandolo sensibilmente dal suo valore normale), hanno posto questo indice in una situazione di utilizzo condizionato molto limitante al punto che nel 2005, Jorge Hirsch, proponendo un nuovo indice -più robusto- che potesse assolvere allo stesso compito [Hir05], catturò l'attenzione dell'intero mondo della ricerca.

La metrica da lui introdotta, l'*h-index*, è il nuovo standard de facto per la valutazione, non solo del singolo ricercatore, ma di tutti i prodotti della ricerca; una possibilità che estende l'adozione dell'indice al di fuori del recinto dei journal di JCR, e dentro al panorama delle conferenze e degli articoli fuori rivista.

Sebbene Hirsch osservi che la sua metrica dia una valutazione equa tra ricercatori, solo se questi abbiano la stessa età, h-index presenta altre problematiche. Ad esempio non fa distinzione tra citazione ed auto-citazione, e nemmeno tra autore e co-autore. Inoltre mette sullo stesso piano di importanza ogni prodotto della ricerca: dalla pubblicazione famosa ma ormai vecchia e citata solo perché "pionieristica", alla ricerca più recente e sperimentale.

A tal proposito, Sidiropoulos, Katsaros e Manolopoulos hanno introdotto un ambiente di indici basato su h-index, che potesse condividerne i pregi di robustezza, migliorandone la sensibilità alle dinamiche citazionali [SKM06]: l'età della citazione, la novità della ricerca, l'indipendenza dall'età del ricercatore.

Questo insieme di indici, assieme all'h-index di Hirch, prende il nome di *Generalized h-index framework*.

h-index : è l'indice introdotto da Hirsch [Hir05], ed è definito secondo la seguente regola:
un ricercatore ha h-index h se h delle sue Np pubblicazioni hanno ricevuto almeno h citazioni l'uno, e le restanti $(Np - h)$ pubblicazioni hanno ricevuto non più di h citazioni

Contemporary h-index (hc-index) : per rendere la valutazione della ricerca sensibile all'età della pubblicazione, e cioè in grado di distinguere la ricerca più datata da quella più *contemporanea*, viene definito un punteggio citazionale $S^c_{(i)}$ per una data pubblicazione i secondo l'equazione

$$S^c_{(i)} = \gamma * (Y_{(now)} - Y_{(i)} + 1)^{\delta} * \mid C_{(i)} \mid$$

dove

Y_{now} è l'anno corrente

Y_i è l'anno della pubblicazione i

$C_{(i)}$ è il numero di pubblicazioni citanti i

e dati $\delta = 4$ e $\gamma = 1$ rispettivamente [SKM06], hc-index è definito secondo la seguente regola:
un ricercatore ha *hc-index* h^c se h^c delle sue Np pubblicazioni hanno ricevuto un punteggio di almeno $S^c_{(i)}$ l'una, secondo l'equazione $S^c_{(i)} \geq h^c$ mentre le restanti $(Np - h^c)$ pubblicazioni hanno ricevuto non più di h^c citazioni ciascuna, rispettando cioè l'equazione $S^c_{(i)} \leq h^c$

Trend h-index (ht-index) : per rendere la valutazione della ricerca sensibile all'età della citazione, e dunque in grado di valutare la novità (il *trend* della ricerca condotta), viene definito un punteggio citazionale $S^t_{(i)}$ per una data pubblicazione i secondo l'equazione

$$S^t_{(i)} = \gamma * \sum_{\forall x \in C_{(i)}} (Y_{(now)} - Y_{(x)} + 1)^{\delta}$$

dove

Y_{now} è l'anno corrente

Y_i è l'anno della pubblicazione i

e dati $\delta = 4$ e $\gamma = 1$ rispettivamente [SKM06], ht-index è definito secondo la seguente regola:
un ricercatore ha *ht-index* h^t se h^t delle sue Np pubblicazioni hanno ricevuto un punteggio di almeno $S^t_{(i)}$ l'una, secondo l'equazione $S^t_{(i)} \geq h^t_{(i)}$ mentre le restanti $(Np - h^t)$ pubblicazioni hanno ricevuto non più di h^t citazioni ciascuna, rispettando cioè l'equazione $S^t_{(i)} \leq h^t$

Normalized h-index (hn-index) : poiché h-index basa il suo calcolo sul numero di pubblicazioni di un ricercatore, e che non tutti i ricercatori

pubblicano lo stesso numero di ricerche nella loro carriera, h-index non risulta molto equa nella valutazione della contribuzione al mondo della ricerca, soprattutto per coloro che abbiano pubblicato meno di altri. Si definisce allora una *normalizzazione* di h-index secondo la seguente definizione:
un ricercatore ha h^n index

$$h^n = \frac{h}{Np}$$

se h delle sue Np pubblicazioni hanno ricevuto almeno h citazioni ciascuna, mentre le restanti $(Np - h)$ pubblicazioni hanno ricevuto non più di h citazioni.

Yearly Conference/Journal h-index (hy-index) : una conferenza, rivista, o journal ha h_y index in un dato anno y, se $N_{p,y}$ delle pubblicazioni uscite durante l'anno y hanno ricevuto almeno h_y citazioni ciascuna, mentre le restanti $(N_{p,y} - h_y)$ hanno ricevuto non più di h_y citazioni.

Normalized Yearly Conference/Journal h-index (hny-index) : come per il caso degli autori o dei gruppi di ricerca, anche una conferenza, una rivista, o un journal pubblicheranno tipicamente un numero variabile di articoli.
Si definisce quindi una normalizzazione dell'indice hy-index:
una conferenza, rivista, o journal ha h_y^n index in un dato anno y

$$h_y^n = \frac{h_y}{N_{p,y}}$$

se h_y delle sue $N_{p,y}$ pubblicazioni nell'anno y hanno ricevuto almeno h_y citazioni ciascuna, mentre le rimanenti $(N_{p,y} - h_y)$ pubblicazioni hanno ricevuto non più di h_y citazioni.

1.2.2 I dataset ed i portali della ricerca

La necessità di tali metriche per la valutazione della ricerca, e la disponibilità di dati sempre più ampia, ha portato all'emergere di nuovi dataset

Open Access, come *GoogleScholar*[7], o *Scopus*[8].
Questi aprono nuove possibilità allo studio del panorama disciplinare della ricerca, e lo fanno attraverso il libero accesso alle fonti citazionali.

Di seguito si illustrano alcuni tra i più noti portali e dataset disponibili alla ricerca.

Thomson Reuters Web of Knowledge (WoK) [9]: sebbene WoK offra accesso alle bibliografie di migliaia di journal, non riporta le citazioni al di fuori della sfera d'appartenenza al JCR[10] (tipicamente journal con una certa costanza di pubblicazione, per i quali l'IF sia facilmente e regolarmente calcolabile), non considera le conferenze, contiene citazioni errate a causa di fenomeni di omonimia e sinonimia di nomi e acronimi, ed è relativamente limitato ad alcune lingue, ed ambiti della ricerca [HM11].

DBLP [11]: il dataset dell'Università Trier, al contrario, cerca di risolvere gli errori di omonimia relativamente ai nomi degli autori e delle publication venue. È inoltre una buona sorgente citazionale per quanto riguarda le conferenze, ma è limitato nel corpo documentale (principalmente alla sfera della Logica e delle Basi di Dati) [Ley09b].
Partendo inoltre come un esperimento accademico, DBLP non ha una qualità di mantenimento al livello di WoK, ed è anzi in procinto di abbandonare la registrazione dei riferimenti citazionali [Ley09b].

GoogleScholar : sempre più usato, soprattutto grazie alla profonda integrazione con il motore di ricerca di Google, GoogleScholar spazia moltissimi ambiti disciplinari, ma soffre di grossi problemi nel conteg-

[7] http://scholar.google.com/
[8] http://www.scopus.com/home.url
[9] http://wokinfo.com
[10] http://thomsonreuters.com/journal-citation-reports/
[11] http://www.informatik.uni-trier.de/ ley/db/

gio delle citazioni ricevute [Lab10], le quali potrebbero essere gonfiate indiscriminatamente, distorcendo la valutazione della ricerca.

Arnetminer DBLP+C [12]: il dataset su cui si basa *ArnetMiner*, un altro portale Open Access, è stato costruito a partire da un esperimento accademico che coniuga DBLP con le citazioni di ACM [Tan+08]. Questi presenta gli stessi vantaggi e limiti di DBLP, ma soffre di qualche problema sul lato delle citazioni, le quali sono integrate a mezzo pattern-matching dal dataset di ACM.

Il mantenimento di DBLP+C condivide lo stesso livello di mantenimento di DBLP, ma resta più duttile: è infatti possibile contattare i suoi mantainer per richiedere correzioni e modifiche, o tenere traccia di paper non associati ad alcuna rivista o journal di pubblicazione.

SCImago Journal/Country Rank (SJR) [13]: dalla collezione delle citazioni provenienti dal dataset *Scopus* di *Elsevier*[14], SJR costruisce ed offre gratuitamente delle statistiche per la valutazione della ricerca, dal più noto IF, al *Scimago Journal Rank*: una metrica definita ad hoc, basata sul conteggio delle citazioni ricevute, ed una misura di "vicinanza" tra journal [BMA12].

Publish or Perish (PoP) [15]: basato sulle citazioni fornite da GoogleScholar, PoP è un programma (installabile su ogni desktop) che espone statistiche atte alla valutazione della qualità e dell'impatto dei prodotti della ricerca.

Data la completa automazione del processo di collezione delle citazioni e la mancanza di integrazione con altri dataset, la qualità dei risultati offerta da PoP non si discosta da quella dei dati offerti da GoogleScholar.

[12] http://arnetminer.org/
[13] http://www.scimagojr.com/
[14] http://www.elsevier.com/
[15] http://www.harzing.com/pop.htm

Eigenfactor [16]: Eigenfactor [WBB10] propone una valutazione dell'intero panorama della ricerca all'interno della sfera di JCR, pesando la rilevanza dei journal citati servendosi dell'algoritmo *Page Rank* [BP98].
Così facendo, le citazioni provenienti da journal autorevoli acquistano maggiore peso rispetto a quelle provenienti da journal meno rilevanti; il rank calcolato dunque risente anche della dimensione del journal in esame.
Oltre ad una metrica, Eigenfactor è un vero e proprio portale per la ricerca, le cui statistiche possono essere consultate on-line liberamente.

Come si è visto, ognuno di questi presenta vantaggi e svantaggi nella propria adozione, per questo, molto spesso, come nel caso di Arnetminer, sono preferiti studi ed esperimenti che prendano in considerazione più di un dataset [MY06; Tan+08; RT05].

Ora che si ha un'idea della varietà dei portali on-line, e delle risorse che questi mettono a disposizione, è necessaria una breve panoramica sulle tecnologie che permettano di utilizzarne i dati citazionali grezzi.
Queste si dividono principalmente in tre metodologie altamente integrabili tra loro, ed il cui uso è indipendente dall'adozione di una tassonomia piuttosto che una collezione di termini folksonomici.

Cluster Analysis : i documenti e le citazioni della rete citazionale in esame vengono visti rispettivamente come *nodi* ed *archi* direzionati di un *grafo* citazionale: un nuovo modello dei dati col quale, mediante opportuni editor (ad esempio *Gephi* [BHJ09]) o librerie grafiche [BOH11; Ach+13], è possibile manipolare e ridisporre nodi ed archi a piacimento, ad esempio ignorando la direzionalità dei link, o anche attribuendovi pesi, o ridisegnandone la forma dei nodi o l'intero layout.
Le possibilità di manipolazione offerte sono virtualmente infinite.
Questo insieme di strumenti e framework grafici sono molto utili per

[16] http://www.eigenfactor.org/

una visualizzazione facilitata, ed una diretta osservazione delle proprietà dei grafi. Sono inoltre sempre più diffusi e di facile prova ed installazione.

L'aspetto presentazionale del grafo, oltre che una mera questione estetica, rappresenta una funzionalità molto potente per mezzo della quale è possibile studiare il grafo attraverso la visualizzazione di fattori comuni. Un esempio pratico è dato dalla effettiva capacità di questi strumenti di raggruppare nodi o archi a seconda del valore di alcuni attributi (ad esempio per data di pubblicazione, o per publication venue).

Machine Learning : un insieme di tecnologie dell'*Intelligenza Artificiale* che permettono di spezzare il corpo testuale di un documento, ed analizzarne i termini con il riferimento al proprio contesto frasale.

Sebbene apparentemente fuori dal contesto dell'analisi di una rete citazionale, queste tecnologie hanno già permesso l'individuazione di trend di ricerca evitando ai ricercatori l'intero processo di costruzione di un dataset delle citazioni [JPJ14].

Lo studio di una rete citazionale potrebbe trovare in queste tecnologie un valido partner da affiancare in fase di test o di indagine congiunta.

Distance Matrix : usata in generale nell'ambito della clusterizzazione di punti in uno spazio euclideo, è stata impiegata anche per attribuire un valore di prossimità ai documenti di uno spazio citazionale, trasformando il numero delle citazioni ricevute in un'informazione di distanza tra punti [LV06; Ley09a; LCR12].

Capitolo 2

Una nuova metrica per la valutazione della ricerca

Nel presente capitolo si propone una nuova metrica per la valutazione della multidisciplinarietà della ricerca. In particolare si dà una definizione di multidisciplinarietà, e si illustrano gli elementi necessari a descrivere un prodotto della ricerca nel suo contesto disciplinare.

Il nuovo indice di valutazione si inserisce nel contesto delle bibliometriche descritte nel capitolo 1, un panorama ricco di indici di valutazione, i cui limiti sono riassunti in due categorie di problemi:

- la mancanza di una semantica della citazione (ad esempio per criticare, o per attribuire merito alla ricerca) può abbassare o aumentare sensibilmente l'effettiva quantificazione della contribuzione

- la mancanza di un contesto disciplinare della citazione che, a parità di contribuzione alla ricerca, comporta indici di impatto maggiore nei ricercatori *monodisciplinari*, e minori ai ricercatori *multidisciplinari* (le cui pubblicazioni non sono al "centro" di una disciplina, ma ai "bordi" di questa, citati cioè da differenti discipline, in un contesto di comportamenti e frequenze citazionali eterogenei)

Il presente lavoro affronta il secondo di questi due problemi, basandosi non sull'adozione delle tassonomie delle discipline viste nel capitolo 1.1.1, bensì procedendo secondo un metodo bottom-up di individuazione degli insiemi disciplinari: sfruttando cioè le proprietà di un grafo citazionale, e proponendo un indice di multidisciplinarietà sensibile al contesto in cui la citazione è attribuita.

2.1 La multidisciplinarietà

In questa sezione si dà una definizione del concetto di multidisciplinarietà di un prodotto della ricerca, e si illustra un metodo di identificazione delle discipline in una rete citazionale.

2.1.1 La metrica multidisciplinare

Definizione 2.1.1.1. *Si definisce* multidisciplinarietà di un prodotto della ricerca *una caratteristica di contribuzione* attribuita *dalla rete citazionale contestuale all'opera considerata. In particolare un'opera è multidisciplinare quando le citazioni che riceve le attribuiscono rilevanza nel contesto di più ambiti disciplinari.*

Secondo la definizione 2.1.1.1, un'opera è multidisciplinare quando presenta due caratteristiche:

rilevanza : le citazioni che riceve ne definiscono il contributo come rilevante all'interno dell'insieme disciplinare citante.
Esistono discipline dove è presente naturalmente un'alta densità di citazioni tra i prodotti della ricerca, e discipline dove la densità citazionale è invece più bassa. La rilevanza di una particolare opera consiste in una soglia citazionale che questa deve presentare per poter essere considerata parte di quella disciplina; una caratteristica indipendente dalla valutazione dell'impatto che la stessa opera esercita all'interno della disciplina citante.

cardinalità disciplinare : le citazioni che riceve provengono da più di una disciplina.

Un documento è considerato multidisciplinare nel momento in cui è rilevante per più di un insieme disciplinare.

L'indice di multidisciplinarietà di un prodotto della ricerca è quindi funzione di queste due caratteristiche: rilevanza, e cardinalità disciplinare.
Si noti che la multidisciplinarietà di un prodotto della ricerca non è definita come una misura dell'impatto che questi eserciti all'interno delle discipline citanti, ma solo come una funzione della cardinalità disciplinare condizionata alla rilevanza attribuitale dalle discipline citanti.

In ultimo, la multidisciplinarietà non è funzione delle citazioni *uscenti* (ovvero dalle voci di bibliografia del particolare articolo valutato), ma funzione delle sole citazioni *entranti*, attribuite dagli altri prodotti della ricerca; un vincolo introdotto con lo scopo di rendere la valutazione più robusta, ed evitare che un autore gonfi il proprio indice di multidisciplinarietà attribuendo citazioni ad altri prodotti della ricerca disciplinarmente distanti rispetto al proprio.

Per poter applicare queste definizioni e valutare la multidisciplinarietà di un prodotto della ricerca, è necessario studiare una rete citazionale che disponga di diverse comunità disciplinari.

2.1.2 Comunità disciplinari e clustering

In questa sezione si illustrano i metodi per l'individuazione delle discipline all'interno di una rete citazionale.

Lo studio del contesto disciplinare di un prodotto della ricerca si svolge collezionando le voci di bibliografia di un insieme di pubblicazioni, e rappresentandone il corpo bibliografico come un *grafo citazionale*. Si ottiene una struttura nella quale:

- ogni pubblicazione è un nodo di attributi *titolo*, *autori*, *data di pubblicazione*, e *publication venue*
- ogni voce di bibliografia è un link citazionale, ovvero un arco tra due nodi: il nodo citante, ed il nodo citato.

In questa nuova struttura si attribuisce inoltre un peso ad ogni arco: un valore di *neighborhood overlap* [EK10], ovvero un indice di *embeddedness* relativo alla rete citazionale che il nodo citante condivide con il nodo citato.
In particolare, per ogni coppia di nodi A e B alle estremità di un link citazionale, la loro neighborhood overlap vale:

$$neighborhood\ overlap_{(A,B)} = \frac{numero\ dei\ nodi\ vicini\ di\ A\ e\ B}{numero\ dei\ nodi\ vicini\ sia\ di\ A\ che\ di\ B}$$

La rete così costruita assume la forma e le proprietà di una rete *Small World* [WS98; Kle00; EK10].

Definizione 2.1.2.1. *Una rete è di tipo Small World se il numero previsto di salti tra una coppia di nodi scelti a caso è piccolo secondo una certa definizione di scala. In particolare, se il cammino minimo medio tra coppie di nodi* D *rispetto al numero totale di nodi* N *nella rete rispetta la seguente funzione:*

$$D = K \log N$$

In questa rete è presente una elevata interconnessione tra nodi affini (le comunità disciplinari omogenee), e bassa interconnessione tra discipline differenti. Quindi

- i documenti monodisciplinari di un dato contesto disciplinare condividono in media più link (le citazioni) da e per i documenti a loro affini
- i documenti multidisciplinari condividono un numero inferiore di link da e per i documenti a loro affini, svolgendo una funzione di "ponte" tra le varie discipline (le comunità, o *cricche*) della rete in esame

Di qui si comprende bene la disparità di un'eventuale valutazione della ricerca che non tenga conto della multidisciplinarietà di un contributo.

Questo approccio bottom-up all'individuazione delle discipline è analogo alla costruzione della rete citazionale effettuata da Leydesdorff [Ley09a; LCR12], ed è motivata dai vantaggi derivanti dall'abbandono dell'uso di una tassonomia delle discipline (come illustrato nel capitolo 1.1.1). Queste sono quindi individuate attraverso l'analisi delle proprietà della rete Small World, usando un algoritmo di *hard clustering* [Sch07], il quale partiziona la rete citazionale in insiemi di nodi tra loro disgiunti.

In questo lavoro si utilizza l'algoritmo di graph-clustering di Chris Biemann *Chinese Whispers* [Bie06], il quale presenta due caratteristiche molto utili allo scopo:

- la possibilità di non dover specificare un numero di cluster a priori;
- il metodo di partizionamento della rete: fortemente dipendente dall'attribuzione di un peso e non direzionalità degli archi.

Sebbene apparentemente controintuitiva, la decisione di non valutare il verso dei link citazionali è dettata principalmente dal fatto che, nonostante l'innegabile direzionalità della citazione, questa è irrilevante ai fini dell'individuazione delle comunità disciplinari della rete. La forza con la quale i nodi sono connessi tra loro viene stabilita dall'embeddedness che questi condividono coi loro vicini (la neighborhood overlap), mentre il verso del loro legame indica solamente una tendenza citazionale, ortogonale alla presenza delle comunità presso cui i nodi sono distribuiti.

La direzione della citazione spiega la tendenza con la quale certi insiemi di autori, gruppi di ricerca, journal, e conferenze si citano tra loro. Ciò è d'aiuto alla comprensione di uno schema citazionale particolare (ad esempio articoli di Fisica citanti articoli di Matematica, e non viceversa), ma non è invece necessaria ai fini della comprensione dei confini delle comunità e della

forza con la quale queste sono legate tra loro (analizzabili invece per mezzo del fattore di embeddedness tra i nodi della rete).

L'applicazione di questo algoritmo di clustering definisce quindi un nuovo modello dei prodotti della ricerca, nel quale ogni pubblicazione è un nodo in un grafo citazionale partizionato in discipline.

2.2 Ridefinire la descrizione dei prodotti della ricerca

Per descrivere i nodi e gli archi della rete citazionale si deve disporre di un insieme di tecnologie e di dettagli di ogni prodotto della ricerca, in grado di definire il tipo di contribuzione, ed il contesto nella quale questa è citata. Questo è possibile attraverso la raccolta e l'organizzazione ragionata dei *metadati* delle pubblicazioni di ogni disciplina.

Di seguito si illustrano le tecnologie e gli elementi necessari alla descrizione dei prodotti della ricerca.

2.2.1 Disambiguazione del tipo di un documento

Un attributo utile alla descrizione di un prodotto della ricerca è il tipo della pubblicazione.
I processi editoriali, e le edizioni di conferenze ed eventi rendono possibili più versioni di una stessa opera, la quale può assumere diverse forme sia nell'aspetto, che nel contenuto.

In questa sezione si introducono le tecnologie necessarie alla rappresentazione degli attributi di una pubblicazione, ed un nuovo paradigma per la formalizzazione dei prodotti della ricerca.

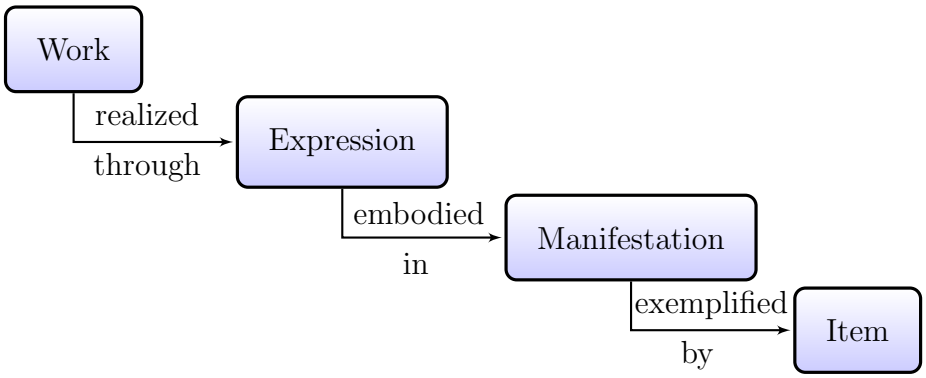

Figura 2.1 Modello di classificazione di un documento secondo IFLA *FRBR*

La descrizione del "tipo della pubblicazione" (e più in generale dell'opera) nei termini del proprio formato, e dei dati contestuali alla propria pubblicazione, si riferisce al modello di classificazione definito dall'*International Federation of Library Associations (IFLA)*[1]: *Functional Requirements for Bibliographic Records (FRBR)* [FRfBRLACSC98], come mostrato in figura 2.1. Questo modello è composto da quattro classi principali:

Work : una particolare creazione intellettuale

Expression : la particolare forma di realizzazione del *Work*, ovvero tutte le versioni, traduzioni, rappresentazioni che, sebbene possano presentare contenuti differenti l'una dall'altra, fanno capo tutte allo stessa opera.

Manifestation : la rappresentazione di una *Expression* sulla base delle caratteristiche del medium che la contiene, come per esempio un libro, o un file pdf.

Item : un esemplare particolare della *Manifestation*, come ogni copia di un libro, oppure ogni copia di un file pdf (identico bit a bit).

Un nuovo paradigma per la descrizione dei prodotti della ricerca consiste nell'attribuire una semantica agli attributi di una pubblicazione.
Questa operazione contemporaneamente logica e linguistica [AH08] è parte

[1] http://www.ifla.org/

di una famiglia di tecnologie e protocolli sviluppati all'interno del *Semantic Web*: un movimento all'interno del *W3C*, atto alla promozione di formati e protocolli standard volti al riuso dei dati in differenti contesti applicativi.

Una più chiara definizione del tipo della pubblicazione e del contesto della contribuzione, si risolve con l'adozione di una famiglia di *ontologie*.

Nel Semantic Web, un'ontologia è la formalizzazione di un sapere concettualizzato: una gerarchia di concetti e loro interrelazioni, le cui proprietà e descrizioni sono esplicitate e formalizzate senza ambiguità.
In ogni ontologia ad esempio, una relazione tra due concetti viene espressa con una funzione, con un proprio dominio e codominio d'azione, una propria cardinalità (le *restrizioni* di una proprietà), ed una propria caratteristica (come la transitività e la riflessività).

Un'ontologia per la descrizione e concettualizzazione di una citazione in questi termini è già disponibile [PS12], ed anzi si inserisce all'interno della famiglia *Semantic Publishing and Referencing Ontologies (SPAR)* [PS12; PSV12].
SPAR nasce in un contesto diverso da quello delle metriche citazionali, per la precisione con il *Semantic Publishing*.
L'esigenza del reperimento di dati facilmente accessibili a script, ed in generale programmi (*agents*) per la ricerca ed il reperimento di documentazione pertinente, ha generato il bisogno di nuove tecnologie semantiche, le quali potessero relazionare dati e ricerche che andassero oltre il verso della sola citazione bibliografica, ma anche nella direzione di un'integrazione più profonda e funzionale alla consultazione immediata.
A questo proposito, Shotton, Portwin, Klyne, e Miles [Sho+09] mettono in evidenza come un'integrazione profonda tra conoscenze appartenenti a prodotti della ricerca differenti, possano, attraverso le tecnologie del Semantic Web, concorrere alla formazione di una base di conoscenza unificata e di immediata consultazione.

Con l'adozione delle ontologie del Semantic Publishing, la descrizione dei prodotti della ricerca è meno opaca, e più strutturata.

2.2.2 Disambiguazione di una publication venue

La raccolta dei metadati dei documenti della rete citazionale termina con la formalizzazione del tipo della publication venue.
In questa sezione si introducono gli accorgimenti necessari ad una loro disambiguazione, e le tecnologie utili alla rappresentazione del loro tipo e dei loro attributi.

Il problema dell'ambiguità dei nomi di una publication venue è un problema noto al mondo della ricerca [Per+08]: la disambiguazione delle stringhe del nome di un evento, un journal, una rivista, è un compito abbastanza complesso, e la creazione di un authority file attendibile dipende fortemente dai dati di partenza utilizzati. I problemi sono, in generale, l'omonimia, la sinonimia, l'adozione di alcuni stili di abbreviazioni.
Nel caso delle conferenze ad esempio, insieme a queste convenzioni si aggiungono gli stili di scrittura (che aggiungono o tolgono spazi e segni di interpunzione), e le edizioni dell'evento, con la conseguente progressione della data o della stagione riportata. Ad esempio, le seguenti stringhe identificano una stessa conferenza:

```
International Semantic Web Conference
International Semantic Web Conference (posters and demos)
ISWC - International Semantic Web Conference
```

Come le ontologie del Semantic Publishing forniscono una maggiore espressività ai prodotti della ricerca visti al capitolo 2.2.1, anche il tipo della publication venue è descritto secondo lo stesso paradigma: in questo caso le ontologie utili sono

- *Semantic Web Conference ontology (SWC)* [DER],
- *Semantic Web for Research Communities (SWRC)* [Sur+05].

Nel presente lavoro, la disambiguazione e l'assegnazione di un tipo delle publication venue è effettuata affiancando la creazione di un authority file all'uso delle ontologie per la descrizione delle venue della ricerca.

Capitolo 3

Casi d'uso: i dataset DBLP e DBLP+C

Nel presente capitolo si illustrano i dettagli dei dataset *DBLP*[1] e *DBLP+C*[2], utilizzati all'interno di questo lavoro per la valutazione della multidisciplinarietà di un prodotto della ricerca.

Già usato per testare il framework h-index generalizzato [SKM06], DBLP è un dataset ricco di informazioni relative a pubblicazioni, conferenze, e proceeding relativi al mondo delle Scienze dell'Informazione.
A questi si integra la rete citazionale del dataset DBLP+C, i cui record sono analoghi a quelli di DBLP, ma la cui rete citazionale è molto più estesa, ed aggiornata [TZY07; Tan+08].

Come accennato nel capitolo precedente, ogni dataset tende ad avere una propria organizzazione interna, ed un proprio modo di formalizzarne i record. Di seguito si illustrano i dettagli di entrambe queste collezioni, specificandone sintassi e semantica.

[1] http://dblp.uni-trier.de/xml/
[2] http://arnetminer.org/citation

3.1 I record del dataset DBLP

I record del dataset DBLP si dividono in due categorie: quelli che seguono la nomenclatura di BibTeX, e che descrivono il tipo di prodotto della ricerca, e quelli creati ad hoc, che esprimono dettagli come casi di omonimia, o forme di pubblicazione della ricerca al di fuori delle possibilità di BibTeX.

Nel dettaglio, ogni record DBLP contiene un attributo *key*, organizzato come uno Unix path-name, e composto da tre parti:

key namespace : concepita per dare una semantica al record DBLP, questa stringa assume diversi valori, in particolare per rappresentare una pubblicazione della ricerca:

- **"conf"** quando il record rappresenta una conferenza o un workshop paper
- **"journals"** quando il record rappresenta un transaction article, una rivista, una newsletter, o un articolo pubblicato in un journal

publication series : è una stringa facoltativa che rappresenta, con un acronimo, o una stringa arbitraria, la particolare journal serie o la rivista della pubblicazione espressa dal record

key ID : è una stringa arbitraria che rappresenta un identificativo per il record descritto.
Solitamente derivata dal nome degli autori e dall'anno di pubblicazione, la stringa è a tutti gli effetti un URN: è non ripudiabile, e resta immutata anche nel caso di errori di ortografia nelle iniziali degli autori.

L'insieme di queste tre parti compone un ID univoco all'interno del dataset, e viene sfruttato per comporre lo URI del record nel momento della sua consultazione on-line.

Per esempio, il record DBLP *journals/jasis/IorioPV11* è rappresentato dal record DBLP al listato 3.1, ed identificato dall'URI:
http://dblp.uni-trier.de/rec/bibtex/journals/jasis/IorioPV11 .

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<!DOCTYPE dblp SYSTEM "documents/dblp/dblp.dtd">
<dblp>
  <article mdate="2011-09-12" key="journals/jasis/IorioPV11">
    <author>Angelo Di Iorio</author>
    <author>Silvio Peroni</author>
    <author>Fabio Vitali</author>
    <title>A Semantic Web approach to everyday overlapping markup.</title>
    <pages>1696-1716</pages>
    <year>2011</year>
    <volume>62</volume>
    <journal>JASIST</journal>
    <number>9</number>
    <ee>http://dx.doi.org/10.1002/asi.21591</ee>
    <url>db/journals/jasis/jasis62.html#IorioPV11</url>
  </article>
  <!-- ... -->
</dblp>
```

Listing 3.1 Un record DBLP identificato dalla *key ID* "journals/jasis/IorioPV11"

A causa della sua polisemia, il key namespace non può garantire da solo una descrizione precisa di un record DBLP. Questa risulta invece attribuibile attraverso l'interpretazione di quegli elementi contenenti le informazioni sugli autori, e le publication venue.

3.1.1 Gli autori

L'elemento *author* di DBLP, a differenza dello stesso elemento BibTeX, contiene il nome per esteso di uno solo degli autori elencati. È quindi comune trovare record DBLP che presentino più elementi author.
La scelta di esprimere il nome di un autore per esteso invece che adottare una forma presentazionale particolare come "*Cognome, Nome*", si deve ad alcune caratteristiche tipicamente culturali: non tutti gli autori hanno un solo nome ed un solo cognome, sono anzi possibili nomi con titoli, nomi occidentali aggiuntivi, o cognomi derivati dal nome di un antenato.

Per evitare quindi che uno stile presentazionale, comune più ad un insieme di discipline che alla totalità dei ricercatori, possa risultare troppo ambiguo ed aggiungere eventuali altri segni di interpunzione nel nome, DBLP adotta la trascrizione del nome per esteso. Nel farlo, DBLP introduce un indice dei ricercatori, ed aggiunge ai suoi record un elemento *www* con queste caratteristiche:

key namespace : la radice dell'attributo *key* dell'elemento *www* deve essere la stringa *homepages*

title : dentro a *www*, l'elemento *title* deve contenere la stringa riservata "*Home Page*".
L'unicità di questa stringa è chiaramente discutibile, tuttavia DBLP adotta la convenzione di far seguire ad ogni titolo di un'opera il carattere ".", e siccome "Home Page" non è seguito dal ".", l'unicità della stringa risulta essere garantita.
Questo elemento è invece omesso nel momento in cui si vuole risolvere un caso di sinonimia: in tal caso il key namespace è comunque mantenuto, mentre al posto dell'elemento *title* si sostituisce l'elemento *crossref*, contenente una stringa corrispondente ad un attributo *key* di qualche altro elemento *www* (a sua volta rappresentante il record di un autore nell'indice dei ricercatori). Ad esempio:

```
<www mdate="2010-11-23" key="homepages/36/2538">
  <crossref>homepages/65/6380</crossref>
</www>
<!-- ... -->
<www mdate="2010-11-23" key="homepages/65/6380">
  <author>Marina de S&aacute; Rebelo</author>
  <author>Marina S. Rebelo</author>
  <title>Home Page</title>
</www>
```

Listing 3.2 L'*alias record* di un ricercatore, ed il suo record principale, contenente due nomi di persona

Rispettati questi vincoli, DBLP prosegue nella descrizione di un ricercatore attraverso i seguenti elementi:

url : la pagina web personale dell'autore

author : il nome per esteso dell'autore nel registro.

Possono essere presenti tanti campi *author* quanti sono i nomi associati a tale autore/autrice. Un esempio di questo è visibile nel listato 3.3, e rappresenta un approccio utile a risolvere casi di sinonimia o di trascrizione abbreviata del nome.

Nel caso invece di omonimia tra due o più autori, DBLP risolve l'ambiguità aggiungendo alcuni numeri al nome originale. Uno di questi casi è illustrato nel listato 3.4.

```
<www mdate="2012-04-25" key="homepages/36/5117">
  <author>Sara Johansson</author>
  <author>Sara Johansson Fernstad</author>
  <title>Home Page</title>
  <url>http://webstaff.itn.liu.se/~sarjo/</url>
</www>
```

Listing 3.3 Il record di un ricercatore con più nomi

```
<www mdate="2012-02-29" key="homepages/85/2046-1">
  <author>Marco Bianchi 0001</author>
  <title>Home Page</title>
</www>
```

Listing 3.4 Il record di un ricercatore che condivide lo stesso nome con un altro ricercatore

note : un campo facoltativo contenente note generiche sulla persona, ad esempio l'affiliazione del ricercatore.

Non esistendo tuttavia una specifica formale relativamente al contenuto di questo elemento, le informazioni al suo interno non sono utilizzabili per arricchire la descrizione del ricercatore attraverso una lettura automatica del contenuto.

Le condizioni da soddisfare per la referenziazione univoca di un ricercatore sono abbastanza fragili, ed in generale molto naïve. Questo non si deve ad un capriccio progettuale, bensì al fatto che il suo design sia stato progettato per la collezione di un numero più contenuto di nomi.

3.1.2 Publication venue e pubblicazioni

La descrizione di un record DBLP prosegue con la formalizzazione della propria publication venue.
Di seguito si riporta un metodo operativo per l'interpretazione degli elementi XML contenenti tali informazioni.

DBLP definisce due possibili *key namespace* come radice del valore dell'attributo *key* associato a ciascun elemento di un record di pubblicazione: "conf" e "journals". Tali stringhe conferiscono una semantica immediata, ma troppo generica per garantirne una descrizione puntuale.

Una descrizione più precisa è comunque inferibile attraverso la combinazione degli elementi BibTeX che compongono il vocabolario XML DBLP:

(in)proceedings : secondo la specifica di BibTeX, per descrivere una pubblicazione come un articolo nei proceedings di una conferenza, l'elemento *inproceedings* deve includere i campi *author*, *title*, *booktitle*, *year*.
Trattandosi di un record rappresentante uno solo degli articoli dei proceedings, DBLP ha formalizzato la relazione *parte-di* con l'aggiunta di un elemento *crossref*, il cui contenuto specifica il valore dell'attributo key dei *proceedings* della conferenza.
I titoli di entrambi i record, così come dell'eventuale serie di cui i proceedings possono far parte (ad esempio *Lecture Notes in Computer Science - LNCS*), vengono invece mantenuti attraverso gli elementi *title* e *booktitle*.
Un esempio di questo caso particolare è illustrato al listato 3.5.

```
<inproceedings mdate="2008-10-20" key="conf/er/Norrie08">
  <author>Moira C. Norrie</author>
```

```
  <title>PIM Meets Web 2.0.</title>
  <year>2008</year>
  <booktitle>ER</booktitle>
  <crossref>conf/er/2008</crossref>
</inproceedings>

<proceedings mdate="2009-04-16" key="conf/er/2008">
  <editor>Qing Li</editor>
  <editor>Stefano Spaccapietra</editor>
  <editor>Eric S. K. Yu</editor>
  <editor>Antoni Oliv&eacute;</editor>
  <title>Conceptual Modeling - ER 2008,
    27th International Conference on Conceptual Modeling,
    Barcelona, Spain, October 20-24, 2008. Proceedings
  </title>
  <volume>5231</volume>
  <year>2008</year>
  <isbn>978-3-540-87876-6</isbn>
  <booktitle>ER</booktitle>
  <series href="db/journals/lncs.html">
    Lecture Notes in Computer Science
  </series>
  <publisher>Springer</publisher>
</proceedings>
```

Listing 3.5 Un record DBLP rappresentante un articolo all'interno di una serie di proceedings.

article : secondo la specifica di BibTeX, per descrivere un articolo appartenente ad un journal o una rivista, *article* deve includere i campi *author*, *title*, *journal*, *year*.

DBLP non formalizza una distinzione tra journal o rivista, ma adotta una particolare combinazione di elementi e referenze interne per modellare il caso di journal appartenenti ad una data serie.

Per riuscire ad esprimere questa relazione al di fuori delle possibilità del vocabolario BibTeX, DBLP prende in prestito gli elementi *proceedings* e *book* (usati intercambiabilmente) dalla loro funzione abituale,

e vi associa un altra semantica: una situazione che mette bene in vista i limiti dell'adozione di BibTeX nell'organizzazione di una biblioteca digitale.

Un esempio di questo caso è illustrato al listato 3.6.

```
<article mdate="2008-04-15" key="journals/jods/HurtadoPW08">
  <author>Carlos A. Hurtado</author>
  <author>Alexandra Poulovassilis</author>
  <author>Peter T. Wood</author>
  <title>Query Relaxation in RDF.</title>
  <!-- ... -->
  <year>2008</year>
  <journal>J. Data Semantics</journal>
  <crossref>journals/jods/2008-10</crossref>
</article>

<proceedings mdate="2008-04-15" key="journals/jods/2008-10">
  <editor>Stefano Spaccapietra</editor>
  <title>Journal on Data Semantics X</title>
  <booktitle>J. Data Semantics</booktitle>
  <series href="db/journals/lncs.html">
    Lecture Notes in Computer Science
  </series>
  <publisher>Springer</publisher>
  <year>2008</year>
  <isbn>978-3-540-77687-1</isbn>
</proceedings>
```

Listing 3.6 Un record DBLP rappresentante un articolo all'interno di una journal series

incollection : secondo la specifica di BibTeX, per descrivere un articolo completo di titolo, ed appartenente ad un particolare volume, *incollection* deve includere i campi *author*, *title*, *booktitle*, e *year*.

DBLP adotta questo elemento quando deve descrivere tutti quegli *article* il cui medium contenitore non sia per forza appartenente ad una

serie. Il suo uso risulta quindi intercambiabile con quello dell'elemento *author*.

editor : le conferenze sono solitamente presiedute da *editor*. Questi possono dividersi in diverse categorie, come i *chair* o *general chair*.
Questo elemento, non solo non è in grado di fornire una distinzione tra i due concetti, ma alle volte assume anche il nome dell'editore di una pubblicazione.
Il team di DBLP sconsiglia di riferirsi a questo elemento per operare qualunque valutazione in merito alla ricerca [Ley09b].

Oltre alle regole definite da DBLP, la presente lavoro si serve di espressioni regolari sul contenuto degli elementi *title*, *booktitle*, e *journal* per documentare anche altri tipi di pubblicazioni, come paper, workshop paper, e poster.

3.2 I record del dataset DBLP+C

Ai record DBLP, il presente lavoro integra la rete citazionale del dataset DBLP+C.

Di seguito si riportano i dettagli di organizzazione dei suoi record, strutturati in una serie di stringhe non vuote disposte su più righe, ed organizzate come una lista di coppie chiave-valore terminate dal carattere di fine riga:

```
#* inizio del campo "titolo"
#@ inizio del campo "autori" (separati tra loro da virgole)
#t inizio del campo "data"
#c inizio del campo "publication venue"
#index inizio del campo "DBLP+C ID della pubblicazione"
#% inizio del campo "DBLP+C ID di una pubblicazione citata"
#% (può occorrere più volte)
#! inizio del campo "abstract della pubblicazione" (facoltativo)
```

3.2.1 Gli autori

Un record DBLP+C mantiene l'ordine degli autori espresso da DBLP, ma li raggruppa in un'unica stringa, separandoli da virgole.
Il presente lavoro sfrutta questa organizzazione per individuare un record DBLP+C che abbia una rappresentazione nel dataset DBLP, ed associarvi la rete citazionale, arricchendo così la descrizione del prodotto della ricerca.

Ai listati 3.7 e 3.8 si mostrano le organizzazioni che i due dataset adottano alla rappresentazione degli autori di un prodotto della ricerca.

```
<inproceedings mdate="2010-04-25" key="conf/sac/MatteoPTV10">
  <author>Nicola Raffaele Di Matteo</author>
  <author>Silvio Peroni</author>
  <author>Fabio Tamburini</author>
  <author>Fabio Vitali</author>
  <title>
    Of mice and terms: clustering algorithms
        on ambiguous terms in folksonomies.
  </title>
  <pages>844-848</pages>
  <year>2010</year>
  <booktitle>SAC</booktitle>
  <crossref>conf/sac/2010</crossref>
</inproceedings>
```

Listing 3.7 Un record DBLP avente una rappresentazione DBLP+C (listato 3.8).

```
#*Of mice and terms: clustering algorithms on ambiguous terms in folksonomies.
#@Nicola Raffaele Di Matteo,Silvio Peroni,Fabio Tamburini,Fabio Vitali
#t2010
#cSAC
#index1384280
#%797277
#%108394
#%21545
#%1317851
```

Listing 3.8 Un record DBLP+C avente una rappresentazione DBLP (listato 3.7).

3.2.2 Publication venue e pubblicazioni

Con la chiave per l'indicazione della publication venue, DBLP+C fornisce una descrizione ancora più generica rispetto a quella già fornita dagli elementi del vocabolario XML DBLP: con questa rappresentazione infatti, DBLP+C mescola i nomi delle conferenze con quelle di workshop, journal, riviste, e proceeding. Ad esempio:

```
#cACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery
#cSIGMOD Conference
#cSIGMOD Workshop, Vol. 1
```

sono voci che indicano che il record di appartenenza è, rispettivamente, un workshop paper presentato alla conferenza "SIGMOD", un paper presentato alla conferenza "SIGMOD", un paper facente parte dei proceedings della conferenza "SIGMOD".

In questo lavoro, l'integrazione tra i due dataset vede quindi tre fasi:

fase 1 raccolta dei dati relativi ai ricercatori elencati nel dataset DBLP:

- risoluzione di omonimie e sinonimie;
- creazione di un'authority file dei ricercatori;

fase 2 raccolta dei dati relativi alle pubblicazioni del dataset DBLP:

- descrizione e disambiguazione del tipo di una publication venue;
- creazione di una base di dati dei prodotti della ricerca completi dell'elenco dei loro autori;

fase 3 integrazione tra dataset DBLP e DBLP+C:

- ricerca di record DBLP+C con stessa lista di autori, titolo e data di pubblicazione di un record DBLP;
- integrazione della rete citazionale;

Capitolo 4

CoData: comunità disciplinari, metadati, bibliometriche

CoData, "Context/Community Data" è un progetto che fornisce una descrizione dei prodotti della ricerca, delle loro relazioni, e del loro contesto. Inoltre offre funzionalità di valutazione della ricerca attraverso il calcolo di metriche di impatto[1] e di multidisciplinarietà.

In questo capitolo si illustrano le componenti che prendono parte alla lettura ed integrazione dei dataset DBLP e DBLP+C, la costruzione di una rete documentale rappresentante il mondo della ricerca, e la costruzione di un grafo citazionale delle pubblicazioni.
Infine, si illustra una tecnica di clustering con la quale individuare le discipline della rete citazionale, e si fornisce un metodo operativo per il calcolo della multidisciplinarietà dei prodotti della ricerca.

[1] CoData implementa l'intero *framework h-index* [Hir05; SKM06] sui prodotti della ricerca, gli autori, e le publication venue

CoData divide la costruzione della rete citazionale in fasi:

data retrieval : lettura dei record dei dataset DBLP e DBLP+C, ed attribuzione di una semantica dei dati

data integration : creazione di una *knowledge base* semantica delle pubblicazioni, contenente le informazioni di entrambi i dataset DBLP e DBLP+C

graph analysis : lettura della knowledge base, e rappresentazione del grafo semantico in un grafo delle citazioni (Small World)

bibliometrics : implementazione di indici bibliometrici di impatto e di multidisciplinarietà per la valutazione della ricerca

4.1 Data retrieval e arricchimento dei metadati

I prodotti della ricerca contenuti e descritti dai dataset DBLP e DBLP+C sono formalizzati secondo specifiche differenti [Tan+08; Ley09b]: il primo adotta un vocabolario XML, mentre il secondo esprime ogni record come una lista di chiavi-valori, separati dal carattere di fine linea.

CoData legge i record di ogni dataset, e vi applica un foglio di stile (*stylesheet*) *XSLT* dipendente dal formato usato per rappresentare i dati. Questo stylesheet associa ad ogni campo del record una rappresentazione semantica del proprio contenuto, formalizzandola nel formato *TURTLE*[2].

Di seguito si introduce il procedimento di attribuzione della semantica ai record: dalla rappresentazione TURTLE, alle ontologie utilizzate per relazionarne i concetti e le interazioni.

[2]http://www.dajobe.org/2003/11/ntriplesplus/

4.1.1 Una famiglia di ontologie per il mondo della ricerca

CoData definisce uno stylesheet per ogni formato dei dati di un dataset. Questo stylesheet trasforma ogni campo di un record in uno *statement*: una frase, composta da:

subject : il nome del campo; può essere il nome dell'elemento XML (nel caso di un record DBLP), o del tag (nel caso di un record DBLP+C), o il concetto che questi esprime

predicate : una proprietà del subject

object : un valore per la proprietà del subject del discorso

Ogni statement asserisce un *fatto* relativo al proprio soggetto: una proprietà del soggetto formalizzata utilizzando vocabolari ed ontologie preferibilmente noti[3]. Per questo motivo, CoData sfrutta le ontologie:

SPAR [PS12; PSV12] una famiglia di ontologie ricca di proprietà adatte alla descrizione delle entità bibliografiche nelle sue parti, relazioni, composizioni, e ruoli dei partecipanti alla pubblicazione.

SWC [DER] e SWRC [Sur+05] indicate per la descrizione degli eventi della ricerca, quali conferenze, workshop, e demo.

CoData trasforma così ogni record di un particolare dataset in un *modello semantico*: un insieme di statement, asserenti fatti sul record in esame.

4.1.2 Il modello semantico

In questa sezione si illustrano alcuni esempi di trasformazione dei record: dal modello dei dati DBLP e DBLP+C, al modello semantico definito sfruttando le ontologie per il semantic publishing ed il mondo della ricerca, introdotte nel capitolo 4.1.1

[3] http://www.w3.org/TR/ld-bp/#VOCABULARIES

CoData avvia la lettura di ogni record, ed in base al suo formato originale, vi applica uno stylesheet XSLT, il quale definisce una trasformazione dal modello dei dati originario al modello semantico (formalizzato in TURTLE). Il record del listato 4.1 viene quindi trasformato nella rappresentazione TURTLE al listato 4.2, dove per brevità d'esposizione sono state omesse le dichiarazioni dei namespace[4]

```
<inproceedings mdate="2010-04-25" key="conf/sac/MatteoPTV10">
  <author>Nicola Raffaele Di Matteo</author>
  <author>Silvio Peroni</author>
  <author>Fabio Tamburini</author>
  <author>Fabio Vitali</author>
  <title>
    Of mice and terms: clustering algorithms
        on ambiguous terms in folksonomies.
  </title>
  <year>2010</year>
  <booktitle>SAC</booktitle>
  <crossref>conf/sac/2010</crossref>
</inproceedings>
```

Listing 4.1 Un record DBLP avente una rappresentazione DBLP+C (listato 3.8).

[4] I namespace TURTLE:

```
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix dblp: <http://dblp.uni-trier.de/> .
@prefix fabio: <http://purl.org/spar/fabio/> .
@prefix foaf: <http://xmlns.com/foaf/0.1/> .
@prefix swc: <http://data.semanticweb.org/ns/swc/ontology#> .
@prefix owl: <http://www.w3.org/2002/07/owl#> .
@prefix core: <http://purl.org/vocab/frbr/core#> .
@prefix xs: <http://www.w3.org/2001/XMLSchema#> .
@prefix dcterms: <http://purl.org/dc/terms/> .
@prefix pro: <http://purl.org/spar/pro/> .
@prefix cito: <http://purl.org/spar/cito/> .
@prefix skos: <http://www.w3.org/2004/02/skos/core#> .
@prefix discipline: <http://bagnacan.web.cs.unibo.it/disciplines/> .
@prefix disciplinescheme: <http://bagnacan.web.cs.unibo.it/disciplineschemes/> .
```

```
# FRBR Work, Expression, e Manifestation della pubblicazione
dblp:conf-sac-MatteoPTV10_w a fabio:Work , foaf:Document ;
  dcterms:title
    "Of mice and terms: clustering algorithms on ambiguous terms
    in folksonomies." .
  core:creator
    dblp:homepages-48-1099 , dblp:homepages-19-7770 ,
    dblp:homepages-80-2302 , dblp:homepages-28-3018 .
dblp:conf-sac-MatteoPTV10 a fabio:Expression , foaf:Document , swc:Paper ;
  core:isRealizationOf dblp:conf-sac-MatteoPTV10_w .
dblp:conf-sac-MatteoPTV10_m a fabio:Manifestation , foaf:Document ;
  fabio:isManifestationOf dblp:conf-sac-MatteoPTV10_w .
  fabio:hasPublicationYear "2010" .
# la publication venue della pubblicazione
dblp:conf-sac-MatteoPTV10 core:partOf dblp:conf-sac-2010 .
# autori della pubblicazione
dblp:homepages-19-7770 foaf:name "Nicola Raffaele Di Matteo" .
dblp:homepages-48-1099 foaf:name "Silvio Peroni" .
dblp:homepages-28-3018 foaf:name "Fabio Tamburini" .
dblp:homepages-80-2302 foaf:name "Fabio Vitali" .
```

Listing 4.2 La pubblicazione rappresentata dal record DBLP del listato 4.1, tradotto da CoData in formato TURTLE.

Il modello semantico di 4.2 non contiene alcuna citazione perché nella sua forma originaria, al listato 4.1, il record DBLP non contiene voci bibliografiche, le quali compaiono invece nel corrispondente record DBLP+C (listato 3.8).

CoData salva questo nuovo modello del record DBLP in un *triple store Virtuoso Open Source (VOS)*[5], ed attende che anche il corrispondente record DBLP+C (contenente le voci bibliografiche, ovvero le citazioni verso altre pubblicazioni) venga letto e trasformato a sua volta in triple semantiche (listato 4.3). Di qui, CoData compie l'arricchimento del record originario, integrando i due modelli in un insieme di statements unico.

[5]http://virtuoso.openlinksw.com/dataspace/doc/dav/wiki/Main/

```
# le rappresentazioni FRBR Work e FRBR Expression della pubblicazione
dblpc:dblpcEntry1384280_w a fabio:Work , foaf:Document ;
  dcterms:title
    "Of mice and terms: clustering algorithms on ambiguous terms
    in folksonomies." .
  core:creator
    dblpc:dblpcEntry1384280-author1 , dblpc:dblpcEntry1384280-author2 ,
    dblpc:dblpcEntry1384280-author3 , dblpc:dblpcEntry1384280-author4 .
dblpc:dblpcEntry1384280 a fabio:Expression , foaf:Document ;
  core:isRealizationOf dblpc:dblpcEntry1384280_w .
# le citazioni della pubblicazione
dblpc:dblpcEntry1384280 cito:cites
    dblpc:dblpcEntry21545 , dblpc:dblpcEntry797277 ,
    dblpc:dblpcEntry108394 , dblpc:dblpcEntry1317851 .
# la publication venue della pubblicazione
dblpc:dblpcEntry1384280 swc:isPartOf dblpc:dblpcEntry1384280-container .
```

Listing 4.3 La pubblicazione rappresentata dal record DBLP+C del listato 3.8, tradotto da CoData in formato TURTLE.

4.2 Integrazione dei dati

Di seguito si illustrano le modalità di reperimento ed integrazione dei record per uno stesso prodotto della ricerca, e di come CoData disambigui il tipo di pubblicazione, ed il tipo di evento ad essa contestuale.

4.2.1 Gli autori

Come illustrato nel capitolo 3.1, un record DBLP che rappresenti una pubblicazione può presentare vari campi autore. Sebbene ogni autore venga rappresentato da un record separato, gli elementi *author* di una pubblicazione non si riferiscono all'autore con il suo identificativo, bensì con il suo nome.

Di seguito si illustra come questi identificativi vengano ricercati, ed utilizzati durante la costruzione di un modello semantico.

CoData effettua una lettura del dataset DBLP per parti: prima della costruzione del modello semantico di ogni record relativo ad una pubblicazione, CoData legge il dataset DBLP alla ricerca di tutti quei record che rappresentano un ricercatore. Come illustrato nel capitolo 3.1.1, tali record possono contenere elementi *crossref* utili alla risoluzione di eventuali *alias record*, che vengono quindi tradotti in un secondo identificativo per il medesimo autore. Il record al listato 3.2 viene quindi trasformato nel record al listato 4.4, e scritto nel triple-store.

```
# record del ricercatore
dblp:homepages-65-6380 a foaf:Agent ;
  foaf:name "Marina de Sá Rebelo" ;
  foaf:name "Marina S. Rebelo" .
# alias record del ricercatore
dblp:homepages-36-2538 owl:sameAs dblp:homepages-65-6380 .
```

Listing 4.4 Il record del ricercatore del listato 3.2, tradotto da CoData in formato TURTLE.

Nel caso di ricercatori omonimi, per i quali la disambiguazione venga risolta aggiungendo una stringa numerica al nome, CoData traduce il record del ricercatore con un modello semantico che includa sia il nome con la stringa numerica, sia senza.

Il mantenimento del nome con la stringa numerica si deve alle scelte di design del dataset DBLP+C: questi costruisce i suoi record usando, come nomi degli autori, quelli riportati all'interno degli elementi *author* di un record DBLP, tra cui -appunto- eventuali nomi con stringhe numeriche.

CoData mantiene questo riferimento in modo da garantire l'integrazione di due record per una stessa pubblicazione. Il record del listato 3.4 viene quindi trasformato nel record al listato 4.5.

```
# record del ricercatore
dblp:homepages-85-2046-1 a foaf:Agent ;
  foaf:name "Marco Bianchi 0001" ;
  foaf:name "Marco Bianchi" .
```

Listing 4.5 Il record del ricercatore del listato 3.4, tradotto da CoData in formato TURTLE.

4.2.2 I tipi di una pubblicazione

CoData associa ai prodotti della ricerca descritti nei record DBLP e DBLP+C i modelli semantici corrispondenti ai record dei ricercatori.

Come visto nell'introduzione, questo avviene associando ad un ricercatore un ruolo di autore per la pubblicazione specificata. Resta tuttavia da definire il tipo della pubblicazione, che può essere un paper, un articolo presentato ad una conferenza, un demo paper, un poster, o un articolo nei proceedings di una conferenza.

Di seguito si documenta il procedimento con il quale CoData provvede all'arricchimento del modello semantico di una pubblicazione, con il tipo della pubblicazione stessa.

elemento BibTeX/DBLP	classificazione CoData
series	una "parte di una serie"
@article	un "articolo"
@incollection	un "articolo"
@mastersthesis	una "tesi di un master"
@phdthesis	una "tesi di dottorato"
@proceedings	elemento polisemico. Nessuna attribuzione semantica
@book	il record è un "libro"
@inproceedings	un "conference poster" se e solo se l'elemento *title* del record contiene la stringa "poster"
	un "demo paper" se e solo se l'elemento *title* del record contiene la stringa "demo"
	un "workshop paper" se e solo se l'elemento *title* del record contiene la stringa "workshop"
	un "conference paper" se e solo se l'elemento *title* del record contiene la stringa "conference" ma non la stringa "workshop"
	un "paper" se e solo se l'elemento *title* del record non contiene match per i casi precedenti

Tabella 4.1 La semantica degli elementi BibTeX/DBLP assegnata da CoData.

Come documentato ai capitoli 3.1 e 3.2, i dataset DBLP e DBLP+C non formalizzano espressamente la tipologia della pubblicazione di un loro record. Questo dettaglio resta comunque inferibile attraverso due informazioni contenute nei record del dataset DBLP: il nome dell'elemento DBLP contenente il record (basato su BibTeXo su una specifica del dataset), o il contenuto degli elementi *title*, *booktitle*, e *journal*.
Per il primo caso CoData si basa sulla semantica del nome BibTeX dell'elemento, mentre nel secondo caso sfrutta una serie di espressioni regolari.
In particolare, la semantica attribuita agli elementi è riassunta nella tabella 4.1.

Tramite queste regole CoData arricchisce la descrizione di un record DBLP o DBLP+C, associandovi nuovi fatti relativi al tipo di pubblicazione.

4.2.3 I tipi di una publication venue

In questa sezione si illustrano i metodi che CoData sfrutta, al fine di arricchire il modello semantico con nuovi fatti relativi al contesto di una pubblicazione: la sua publication venue.

L'aggiunta di informazioni sulla publication venue di un prodotto della ricerca prende due percorsi principali: nel primo le nuove informazioni sono inferite dagli elementi del record, in particolare secondo la semantica adottata da BibTeX e dal dataset DBLP. Nel secondo invece, nuovi dettagli sono associati al modello semantico risultante attraverso la costruzione di un authority file delle publication venue.

Per quanto riguarda la semantica degli elementi di un record, come visto nel capitolo 3.1, DBLP associa diversi significati all'elemento BibTeX *@proceedings*.
CoData opera allora una disambiguazione, classificando il record come:

conference proceedings nel contesto di una conferenza : CoData attribuisce il tipo "conference proceedings" al record, e crea un modello per la conferenza contestuale alla pubblicazione

journal nel contesto di una serie : CoData attribuisce il tipo "journal" se e solo se è presente un elemento *journal* contenente un riferimento ad un altro record interno a DBLP.
Parallelamente, CoData verifica la presenza dell'elemento *series*, caso per il quale crea un nuovo URI, al quale attribuisce il tipo di "serie" contestuale alla pubblicazione del record in esame.

Una volta attribuito un tipo, si procede alla disambiguazione delle *stringhe* dei prodotti della ricerca.
Nel mondo della ricerca esistono convenzioni che portano stessi eventi o stesse pubblicazioni ad assumere nomi (stringhe) molto diversi tra loro. Ad esempio il listato 4.6 contiene label riferite alla conferenza *I-SEMANTICS*: alcune

con variazioni del nome, altre con un riferimento alla particolare conference track.

```
I-SEMANTICS
I-SEMANTICS (Posters & demos)
I-SEMANTICS (Posters & Demos)
I-SEMANTICS (posters and demos)
I SEMANTICS
```

Listing 4.6 Differenti label per la stessa conference I-SEMANTICS

CoData risolve omonimia e sinonimia delle publication venue con la creazione di un authority file.
Le stringhe al listato 4.6 vengono così disambiguate nell'authority record al listato 4.7.

```
<?xml version='1.0' encoding='UTF-8'?>
<entities>
  <entity>
    <before>I-SEMANTICS</before>
    <before>I-SEMANTICS (Posters & demos)</before>
    <before>I-SEMANTICS (Posters & Demos)</before>
    <before>I-SEMANTICS (posters and demos)</before>
    <before>I SEMANTICS</before>
    <after>I-SEMANTICS</after>
    <id>3fee71204edf3021c4134fdcf65f524c4ceb467</id>
  </entity>
  ...
</entities>
```

Listing 4.7 Integrazione delle label per la conferenza I-SEMANTICS in un authority file. Agli alias di una stessa conferenza (*before label*) sono associati una label preferenziale (*after label*), ed un hash (*id*)

In conclusione CoData effettua l'ultimo passo di traduzione dai record DBLP/DBLP+C al loro modello semantico: l'integrazione tra modelli di uno stesso prodotto della ricerca.
In questo processo, il record al listato 4.1 diventa il record al listato 4.8

```
dblp:conf-sac-MatteoPTV10
```

```
  a fabio:Expression , foaf:Document , swc:Paper ;
  core:partOf dblp:conf-sac-2010 .
# la conferenza
dblp:conference-b344233b33b16ff3e68077a6fcf5173ad53d91
  a swc:AcademicEvent ;
  rdfs:label "SAC 2010" , "SAC" ;
  swc:hasRelatedArtefact dblp:conf-sac-2010 .
# i proceedings della conferenza
dblp:conf-sac-2010
  a fabioExpressionCollection , fabio:AcademicProceedings , swc:Proceedings ;
  rdfs:label
    "SAC Proceedings of the 2010 ACM Symposium on Applied Computing (SAC),
       Sierre, Switzerland, March 22-26, 2010" .
```

Listing 4.8 Il record DBLP al listato 4.1 arricchito con la descrizione disambiguata della publication venue alla quale appartiene.

Di seguito si illustrano i passi che CoData compie nell'unificazione ed arricchimento dei modelli semantici:

disambiguazione DBLP - publication venue : CoData avvia una lettura del dataset DBLP, e colleziona ogni label di ogni publication venue (procedimento visto nel capitolo 4.2.3). Ogni label viene disambiguata in un file XML che le raggruppa, eleggendo una label preferenziale, ed associandovi un ID univoco all'interno del file (e dipendente dal tipo di publication venue).

trasformazione record DBLP+C : CoData avvia la lettura del dataset DBLP+C, e genera un modello semantico per ogni record (procedimento visto nel capitolo 4.1.2)

disambiguazione DBLP - ricercatori : si avvia una seconda lettura del dataset DBLP, e si genera un modello semantico per ogni ricercatore (come visto nel capitolo 4.2.1)

trasformazione record DBLP : si avvia una terza lettura del dataset DBLP, e si genera un modello semantico per ogni record rappresentante

un prodotto della ricerca (come visto nel capitolo 4.1.2).
Parallelamente CoData verifica se nel triple-store siano già presenti modelli che descrivano gli elementi del record, in tal caso:

- si integrano i modelli degli autori
- si integrano le label che rappresentino la stessa publication venue del record e, nel caso in cui DBLP non specifichi un identificativo per la venue in esame, si usa il nuovo ID associato
- si cercano modelli di record DBLP+C che condividano stessi autori, stessa data di pubblicazione, e stesso titolo. Nel caso ve ne siano, CoData ne integra la rete citazionale

CoData permette una descrizione estesa ed accurata dei prodotti della ricerca: con questo nuovo modello semantico è possibile osservare una rete citazionale ed il contesto della ricerca con il quale questi interagiscono.

4.3 Il clustering delle discipline

Di seguito si introduce il partizionamento della rete citazionale in discipline della ricerca, e si illustra come CoData assegni un ambito disciplinare ad ogni documento del triple-store, senza che questi sia attribuito a priori per mezzo di una tassonomia, ma sfruttando un algoritmo di clustering.

CoData raccoglie dal triple-store tutte le pubblicazioni che citano e che sono citate, e costruisce il *Connected Component* del grafo citazionale: una rete in cui ogni coppia di nodi è collegata da un cammino di archi citazionali.

Un esempio di questa rimodellazione può essere osservata in figura 4.1, dove è rappresentata una porzione del Connected Component incluso nel triple-store.

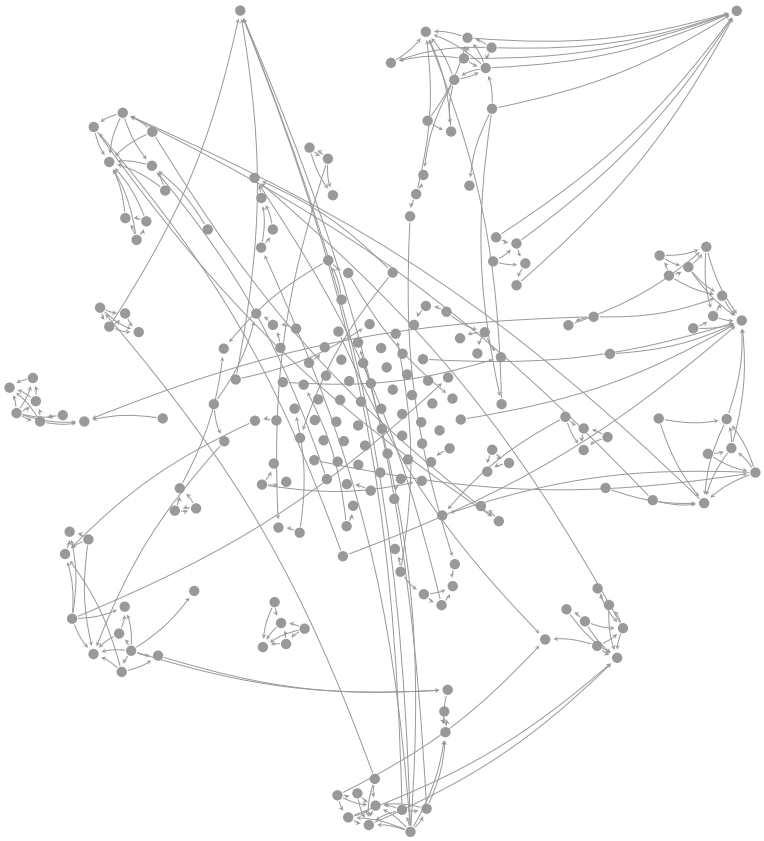

Figura 4.1 Una porzione del Connected Component nel grafo citazionale

CoData procede quindi alla clusterizzazione del Connected Component. Il risultato di tale operazione è riassunto in tabella 4.2, dove è mostrato il numero di cluster della rete citazionale ed il loro contenuto in termini di numero dei documenti.

numero totale di pubblicazioni	380˙527
numero totale di cluster	56˙157
cluster contenenti 1˙000 o più pubblicazioni	27 ($\sim$ 0,007%)
cluster contenenti 100 $\div$ 999 pubblicazioni	358 ($\sim$ 0,09%)
cluster contenenti 10 $\div$ 99 pubblicazioni	3˙819 ($\sim$ 1%)
cluster contenenti 2 $\div$ 9 pubblicazioni	363˙639 ($\sim$ 95,6%)
cluster contenenti 1 pubblicazione	12˙684 ($\sim$ 3,3%)

Tabella 4.2 Il risultato del clustering del Connected Component.

Il risultato alla tabella 4.2 è ottenuto attraverso due passi di elaborazione:

- la rappresentazione di ogni documento con una nuova struttura dati, *DocumentNode*. Questa consiste nella lettura di ogni nodo del Connected Component, e nella raccolta di tutti gli attributi utili alla valutazione del contesto della pubblicazione, come:

  title : il titolo della pubblicazione

  type : il tipo della pubblicazione

  bookSeries : l'identificativo della serie della pubblicazione

  conference : l'identificativo della conferenza d'appartenenza

  proceedings : l'identificativo dei proceedings d'appartenenza

  publisher : l'identificativo dell'editore

  journal : l'identificativo del journal d'appartenenza

  school : l'identificativo dell'ente presso la quale è stata svolta

  citedDocuments : la lista di documenti citati (le voci di bibliografia)

- l'applicazione di un algoritmo di clustering.
  CoData individua le discipline del grafo citazionale clusterizzando il Connected Component per mezzo dell'algoritmo *Chinese Whispers* [Bie06] (implementato nella libreria *S-Space* [JS10]): un algoritmo di hard clustering che agisce su grafi pesati non direzionati.

Ad ogni nodo (documento) del Connected Component viene dunque assegnata una classe (disciplina) d'appartenenza.
Un esempio di questa associazione è al listato 4.9, il quale è arricchito con la descrizione della disciplina d'appartenenza risultante dall'operazione di clustering.

```
# la rappresentazione FRBR Expression della pubblicazione
dblp:journals-tse-CiancariniTVRK98
  a foaf:Document , fabio:Article , fabio:Expression ;
  owl:sameAs dblpc:dblpcEntry1127831 .
```

```
  core:partOf dblp:journal-b14cd4f6ff5a9e1c8ac5976af1449fa6e0a054ef . # venue
# la disciplina della pubblicazione
dblpc:dblpcEntry1127831
  fabio:hasDiscipline discipline:discipline334515 .
discipline:discipline334515
  a fabio:SubjectDiscipline ;
  rdfs:label "discipline334515" .
  skos:inScheme disciplinescheme8db7d .
```

Listing 4.9 Una pubblicazione nel triple-store, arricchita con le informazioni sulla disciplina d'appartenenza.

Lo stesso esempio è invece visibile in figura 4.2, dove è rappresentata la rete citazionale disciplinare contestuale alla pubblicazione del listato 4.9

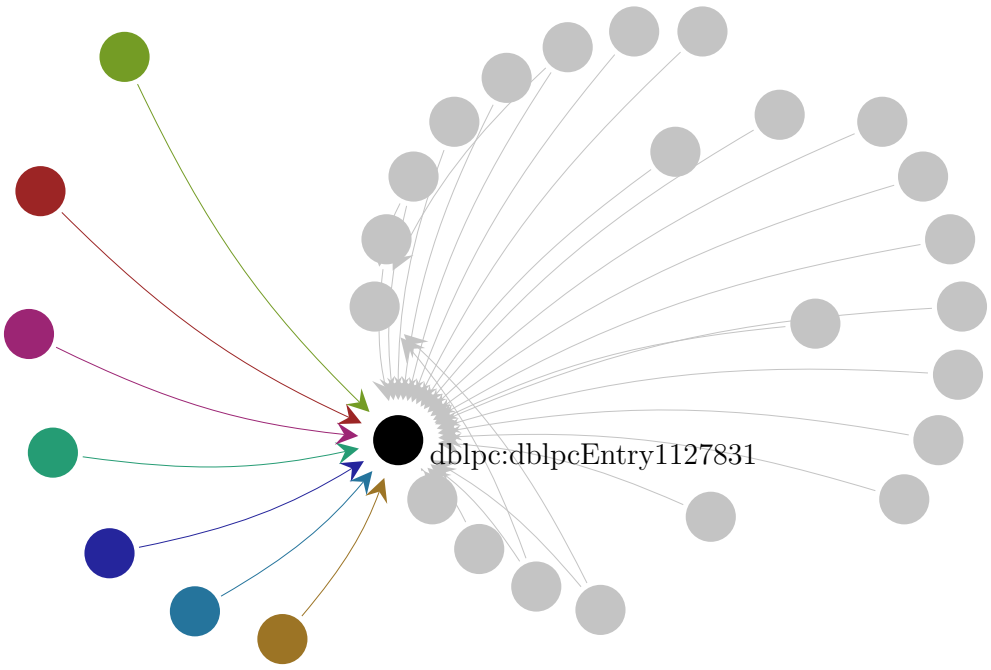

Figura 4.2 La pubblicazione "Coordinating Multiagent Aplications on the WWW: A Reference Architecture." (identificativo dblpc:dblpcEntry1127831) nella sua rete citazionale clusterizzata in discipline

Le classi disciplinari individuate dall'algoritmo di clustering vengono quindi associate ai relativi nodi (le pubblicazioni), e scritte direttamente nel triple-store.

4.4 La multidisciplinarietà

Ad oggi, la valutazione dei prodotti della ricerca si basa sul conteggio delle citazioni che questi ricevono da altre pubblicazioni.
Il panorama della ricerca è però variegato. Una valutazione che tenga conto della diversità disciplinare che un particolare studio mette in relazione, è possibile solamente osservando il contesto disciplinare presso il quale questi è citato.

Di seguito si applica la proposta di multidisciplinarietà secondo la definizione 2.1.1.1, e si illustra una valutazione della stessa per alcuni esempi di prodotti della ricerca.

Secondo la definizione 2.1.1.1, la multidisciplinarietà di un prodotto della ricerca dipende dalle caratteristiche di *rilevanza* e *cardinalità disciplinare* del contributo, mentre è invece indipendente dall'impatto che questa ottiene nella disciplina citante.
Un indice di multidisciplinarietà costruito nel rispetto di questa definizione, è quindi sensibile alla rilevanza di un opera nel suo contesto multidisciplinare.
Le motivazioni di tali vincoli sono riassumibili come segue:

- l'indice di multidisciplinarietà deve potersi inserire nel contesto delle bibliometriche attuali, integrandosi con quelle esistenti, in particolare accompagnando la valutazione dell'impatto di un prodotto della ricerca con quello della multidisciplinarietà della stessa;
- la valutazione della multidisciplinarietà della ricerca deve considerare la centralità del contributo dell'opera rispetto alle comunità disciplinari della rete citazionale, ovvero la capacità dell'opera di collegare tra loro comunità disciplinari tra loro distinte.
- l'indice di multidisciplinarietà deve giudicare rilevante, rispetto ad una disciplina, quei prodotti della ricerca le cui citazioni ricevute siano in accordo con il comportamento citazionale della disciplina citante.

CoData definisce allora la soglia di *rilevanza* per un prodotto della ricerca. Questo valore è funzione del comportamento citazionale interno alla disciplina citante, ed è quindi calcolato osservando la distribuzione delle citazioni tra ogni coppia di discipline[6]: quella citata, contenente l'opera in esame, e quella citante.

Di fronte a questa caratterizzazione dell'indice di multidisciplinarietà M di un prodotto della ricerca x, il valore della multidisciplinarietà $M_{(x)}$ per una data pubblicazione deve:

- essere maggiore di 1: ogni prodotto della ricerca è assegnato ad una disciplina, ne consegue che, se $M_{(x)} = 1$, allora il contributo è disciplinato in una sola disciplina;

- contare il numero di discipline per le quali il contributo risulti rilevante. Una pubblicazione è dunque multidisciplinare quando, oltre alla disciplina di appartenenza, questa è rilevante (ovvero ottiene un numero di citazioni sufficiente) per almeno un'altra disciplina.

Dati questi vincoli, le pubblicazioni x e y citate come in figura 4.3 condividono lo stesso valore di multidisciplinarietà se e solo se entrambe superano le soglie di rilevanza per lo stesso numero di discipline.
Lo stesso può essere detto delle pubblicazioni z e w di figura 4.3, le quali, nonostante non condividano lo stesso numero di citazioni, risultano ugualmente multidisciplinari perché superano le soglie di rilevanza nello stesso numero di discipline.

Diversamente, se la distribuzione citazionale di una disciplina citante con la disciplina citata individuasse una soglia di rilevanza molto alta, ed altre ne individuassero una molto bassa (ad esempio 2 citazioni), i risultati della multidisciplinarietà delle pubblicazioni in figura 4.3 sarebbero molto diversi tra loro.

[6]potrebbe essere utile escludere dal calcolo tutte quelle discipline il cui numero di documenti sia sotto una soglia impostabile dall'utente

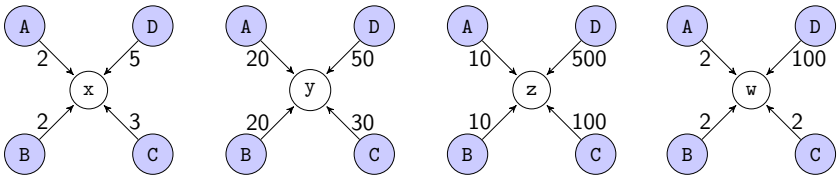

Figura 4.3 Scenari di calcolo della multidisciplinarietà di un prodotto della ricerca x, y, z, w, nel contesto delle citazioni ricevute dalle discipline A, B, C, D.

Allo scopo di osservare la distribuzione delle citazioni tra discipline e calcolare le soglie di rilevanza, CoData costruisce una matrice delle citazioni $n * m$, dove n è il numero di documenti, ed m è il numero delle discipline. Ogni riga è dunque un vettore citazionale di cardinalità m, rappresentante il numero di citazioni che un documento riceve da ogni disciplina.
La matrice ottenuta è sparsissima, ciò significa -per righe- che le pubblicazioni multidisciplinari sono numericamente inferiori alle pubblicazioni monodisciplinari, e -per colonne- che le discipline citanti citino un numero ristretto di pubblicazioni, risultando quindi poco connesse con le altre discipline.

disciplina citata	disciplina citante	1° decile	2° decile	...	10° decile
332895	332895	0.0	1.0	...	149.0
332895	322675	0.0	0.0	...	1.0
332895	274696	0.0	0.0	...	1.0
332895	87219	0.0	0.0	...	2.0
...	...	...	...	...	...
322675	322675	1.0	1.0	...	379.0
...	...	...	...	...	...
274696	274696	1.0	1.0	...	125.0
...	...	...	...	...	...

Tabella 4.3 La distribuzione delle citazioni espressa in decili, tra una disciplina citata ed una disciplina citante

Tale matrice delle citazioni è quindi valutata per parti: si considerano i soli r documenti appartenenti ad una disciplina, e si costruisce una matrice

sparsa $r * m$, dove le r righe sono i documenti della disciplina citata, mentre m sono le discipline citanti.

Riprendendo la definizione 2.1.1.1, la multidisciplinarietà di un prodotto della ricerca è valutata nel contesto della distribuzione delle citazioni tra due insiemi disciplinari: la disciplina d'appartenenza (quella citata), ed ogni altra disciplina tra le m selezionate dalla soglia iniziale (le discipline citanti).
Per avere un'idea dell'andamento citazionale tra ogni coppia di discipline, si procede quindi al calcolo dei percentili della distribuzione delle citazioni[7].
Il risultato di questo calcolo è riassunto in tabella 4.3.

Il valore di multidisciplinarietà M per ogni prodotto della ricerca appartenente alla disciplina citata, è calcolato quindi assegnando un "+1" nel caso in cui il numero di citazioni provenienti dalla disciplina citante superi la soglia di rilevanza definita al 10° decile della distribuzione per la coppia "disciplina citata / disciplina citante".

Con riferimento alla tabella 4.3, se ad esempio una pubblicazione appar-

[7]data m la lunghezza del *feature vector*, il calcolo della distribuzione citazionale avviene mediante la seguente definizione di percentile:

```
if (m = 1) then
  Percentile(p) = l'unico elemento del feature vector
else
  pos = p * (m + 1) / 100 # la posizione percentile
  d = decimal(pos) # la parte decimale di pos
if (pos < 1) then
  Percentile(p) = min({m}) # l'elemento più piccolo nel feature vector
else if (pos >= m) then
  Percentile(p) = max({m}) # l'elemento più grande nel feature vector
else
  lower = integer(pos) # l'elemento del feature vector alla posizione pos
  upper = pos + 1 # l'elemento nel feature vector alla posizione (pos + 1)
  Percentile(p) = lower + (d * (upper - lower))
```

tenente alla disciplina *332895* ottiene una citazione dalla disciplina *322675*, ed una citazione dalla disciplina *87219*, questa non viene considerata multidisciplinare, perché il numero delle citazioni ricevute non supera la soglia di rilevanza stabilita dall'ultimo decile per le coppie (disciplina 332895, disciplina 322675) e (disciplina 332895 , disciplina 87219).

titolo della pubblicazione	id. pubblicazione	M
"The Interdisciplinary Study of Coordination."	dblpc:dblpcEntry832821	14
"Coordinating Multiagent Aplications on the WWW: A Reference Architecture."	dblpc:dblpcEntry1127831	3
"HYTECH: A Model Checker for Hybrid Systems."	dblpc:dblpcEntry79505	3
"Experience Prototyping."	dblpc:dblpcEntry5557	1
"Reachability Analysis via Face Lifting."	dblpc:dblpcEntry257571	1

Tabella 4.4 Un esempio di valutazione della multidisciplinarietà per cinque pubblicazioni nel triple-store

Più in generale, tornando alla figura 4.3, se un documento x appartenente alla disciplina E supera la soglia di rilevanza nelle N discipline, la sua multidisciplinarietà vale $M_{(x)} = n$.
Ad esempio, un documento che supera la soglia di rilevanza nella disciplina D, ma non in A, B, e C, ottiene una multidisciplinarietà $M_{(x)} = 2$.
Se invece lo stesso non supera la soglia di rilevanza in nessuna altra disciplina , allora la sua multidisciplinarietà non è qualificabile: x è valutato come "monodisciplinare", ovvero disciplinato nella sola disciplina E, quella d'appartenenza.

Un esempio del calcolo della multidisciplinarietà dei prodotti della ricerca presenti nel triple-store di CoData è visibile in tabella 4.4. Qui, differenti pubblicazioni in differenti discipline mostrano valori di multidisciplinarietà molto diversi gli uni dagli altri.
Si osserva come la pubblicazione "The Interdisciplinary Study of Coordination" (identificativo dblpc:dblpcEntry832821), dichiaratamente multidisci-

plinare, riceva una valutazione maggiore rispetto alle pubblicazioni più "di settore".

Tra i risultati di tabella 4.4 è ripreso anche l'articolo "Coordinating Multiagent Aplications on the WWW: A Reference Architecture." (identificativo dblpc:dblpcEntry1127831) visto al listato 4.9.
La sua rete citazionale è mostrata in figura 4.2, mentre in figura 4.4 sono rappresentati i contesti disciplinari che citano tale pubblicazione, le cui distribuzioni citazionali si riassumono in tabella 4.5.
Da queste risulta che la pubblicazione in esame supera la soglia di rilevanza in tre discipline diverse dalla propria. Il valore della multidisciplinarietà del contributo vale dunque $M_{(dblpc:dblpcEntry1127831)} = 3$.

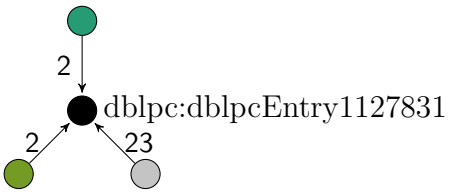

Figura 4.4 Il contesto multidisciplinare e le citazioni ricevute dalla pubblicazione dblpc:dblpcEntry1127831 (la cui rete citazionale è rappresentata in figura 4.2).

disciplina della pubblicazione dblpc:dblpcEntry1127831	disciplina citante	1° decile	2° decile	...	10° decile
334515	116337	0.0	0.0	...	1.0
334515	111060	0.0	0.0	...	1.0
334515	136027	0.0	0.0	...	1.0

Tabella 4.5 La distribuzione delle citazioni espressa in decili tra la disciplina della pubblicazione dblpc:dblpcEntry1127831 citata, e le sue discipline citanti

In questo capitolo sono state illustrate le componenti di CoData, dalla lettura ed integrazione dei dataset DBLP e DBLP+C attraverso un modello semantico, alla valutazione della multidisciplinarietà dei prodotti della ricerca, sfruttando la clusterizzazione di un nuovo modello dei dati: il grafo citazionale.

Capitolo 5

Valutazione

CoData sfrutta diverse tecnologie: la parte di data retrieval e arricchimento si basa su Java, XSLT, Virtuoso Open Source[1], e SPARQL[2], mentre la creazione dell'authority file è implementata in Python.

In questo capitolo si illustrano i tempi di computazione di CoData nelle operazioni di trasformazione dei dataset, clustering del grafo citazionale, e analisi bibliometrica, e si fornisce una valutazione della qualità dei risultati ottenuti.
Quando non specificato, i risultati sono da intendersi ottenuti su una macchina Intel Core 2 Duo 2.4GHz, 8GB RAM DDR3.

5.1 Dai dataset al Linked Data

Come documentato nel capitolo 4.2, la rete citazionale utilizzata per la valutazione della ricerca si avvale di un modello semantico della stessa, ottenuto attraverso la trasformazione dei record di due dataset (DBLP e DBLP+C), e la pubblicazione (e linking) dei loro contenuti in un triple-store.

[1] http://virtuoso.openlinksw.com/dataspace/doc/dav/wiki/Main/
[2] http://www.w3.org/TR/rdf-sparql-query/

La fase di trasformazione dei dataset, dal loro formato originale ad un modello semantico, comprende le operazioni di:

- parsing di ogni record
- trasformazione del record in un modello semantico
- eventuale integrazione del modello semantico con nuovi statements (ad esempio attraverso l'interrogazione di un authority file)
- scrittura del modello nel triple-store

Di seguito si riportano i tempi di esecuzione necessari dall'operazione di recupero dei record in un dataset, alla pubblicazione delle loro informazioni in un triple-store. Successivamente, si procede con una valutazione della qualità dei risultati ottenuti.

5.1.1 La trasformazione di DBLP+C

Come già illustrato nel capitolo 3.2, i record del dataset DBLP+C consistono di una lista di stringhe non vuote, organizzate in coppie chiave-valore, e terminate dal carattere di fine riga (esempio al listato 5.1).

```
#*Coordinating Multiagent Aplications on the WWW: A Reference Architecture.
#@Paolo Ciancarini,Robert Tolksdorf,Fabio Vitali,Davide Rossi,Andreas Knoche
#t1998
#cIEEE Trans. Software Eng.
#index1127831
#%793556
#%832400
#%793734
```

Listing 5.1 Un record nel dataset DBLP+C.

Di fronte a questa organizzazione dell'informazione, il parser di CoData legge il dataset una riga alla volta fino al carattere di teminazione del record, e traduce le coppie chiave-valore in uno statement semantico in un'unica operazione di trasformazione.

Come documentato nel capitolo 4.2, CoData non ha bisogno di arricchire il modello semantico del record in esame, e provvede dunque a scriverlo direttamente in un triple-store dedicato.

Il tempo medio di traduzione di un record DBLP+C è di circa 3 secondi. Il risultato è ottenuto prendendo la media di dieci prove di esecuzione, e calcolato considerando il numero medio (arrotondato per eccesso) di autori per record (3), ed il numero medio di citazioni per record (2).

Qualitativamente, il procedimento di rappresentazione dei record DBLP+C, dalla loro forma originaria al loro corrispondente modello semantico non aggiunge dettagli significativi sulla pubblicazione: questo passo di traduzione è a tutti gli effetti una riscrittura dei tag DBLP+C in una struttura composita, allineata alla famiglia di ontologie SPAR.

L'unico punto di pregio di questa rimodellazione è la pubblicazione dei record DBLP+C nel Linked Data, che li rende disponibili in un triple-store accessibile tramite uno SPARQL end-point.
Il problema di questi dati resta però nella loro qualità: come visto nel capitolo 3.2, DBLP+C non specifica il tipo di publication venue.
Questa traduzione dalla scrittura dei record originaria al Linked Data è apprezzabile solo con il passo di rimodellazione successivo: la trasformazione dei record DBLP.

5.1.2 La trasformazione di DBLP

Come illustrato nel capitolo 3.1, i record del dataset DBLP sono rappresentati da un albero XML la cui semantica degli elementi segue in parte la specifica di BibTeX, ed in parte alcune regole definite dal gruppo di mantenimento di DBLP (esempio nel listato 5.2).

```
<article mdate="2011-10-26" key="journals/tse/CiancariniTVRK98">
  <author>Paolo Ciancarini</author>
  <author>Robert Tolksdorf</author>
  <author>Fabio Vitali</author>
  <author>Davide Rossi</author>
  <author>Andreas Knoche</author>
  <title>
    Coordinating Multiagent Aplications on
    the WWW: A Reference Architecture.
  </title>
  <year>1998</year>
  <journal>IEEE Trans. Software Eng.</journal>
</article>
```

Listing 5.2 Un record nel dataset DBLP.

Diversamente dai record DBLP+C, ai record DBLP sono integrati nuovi statement relativi alla descrizione degli autori, le descrizioni delle publication venue, e la rete citazionale del record DBLP+C corrispondente.

Di fronte a questa organizzazione dell'informazione, il parser di CoData legge il dataset DBLP tre volte: la prima per costruire un authority file delle publication venue, la seconda per collezionare una descrizione degli autori, la terza per leggere ogni record di ogni prodotto della ricerca, ed arricchirlo con i dettagli relativi agli autori, le publication venue, e la rete citazionale di DBLP+C.

Il tempo medio di traduzione di un record DBLP rappresentante un autore è di 0.003 secondi, ed è calcolato considerando il numero medio di alias che lo stesso record contiene.

Per un record DBLP rappresentante una pubblicazione invece, il tempo medio di traduzione ed arricchimento è variabile.
Per questo caso, il calcolo è stato effettuato considerando il tipo di pubblicazione più frequente all'interno del dataset, l'"articolo nei proceedings di una

conferenza"[3]. Questo tipo di pubblicazione presenta in media tre autori, due label da associare alla propria publication venue, e una citazione.
Il tempo di lettura, traduzione, ed arricchimento di un record con queste caratteristiche è circa 6 secondi.

Qualitativamente, la rappresentazione dei record DBLP successiva alla loro formalizzazione originaria, è visibilmente più dettagliata, ed aggiunge particolari significativi sia sulla pubblicazione che sulla struttura del mondo della ricerca ad essa contestuale.

L'integrazione dei dettagli della pubblicazione con il modello semantico degli autori, la publication venue, e la rete citazionale ricavata dal dataset DBLP+C, portano all'ottenimento di un modello semantico molto più chiaro e dettagliato del semplice record taggato con stringhe arbitrarie.

Ai vantaggi di una più puntuale rappresentazione dei dettagli di un prodotto della ricerca, si aggiunge quella della disponibilità degli stessi nel Linked Data. I record DBLP e DBLP+C sono infatti riusabili, e disponibili in un triple-store accessibile mediante uno SPARQL end-point.

[3]i tipi di pubblicazioni più frequenti all'interno del dataset DBLP:

tipo di pubblicazione	numero
articolo nei proceedings di una conferenza	913'759
paper	770'283
proceedings di una conferenza	19'546
libro	9'054
tesi di dottorato	6'913
paper di una demo	1'504
paper di un workshop	1'361
poster di una conferenza	627
paper di una conferenza	545
tesi di un master	9

Un esempio del risultato di rimodellazione è visibile in figura 5.1, dove è rappresentata la struttura di un modello semantico relativo al tipo di pubblicazione "articolo nei proceedings di una conferenza".

5.2 Dalle citazioni alla multidisciplinarietà

Come visto ai capitoli 4.3 e 4.4, la valutazione della multidisciplinarietà di un prodotto della ricerca è ottenuta:

- individuando le comunità disciplinari (clusterizzando il Connected Component attraverso l'algoritmo *Chinese Whispers*);
- analizzando la distribuzione delle citazioni tra discipline differenti (calcolando la soglia di rilevanza).

Di seguito si illustrano i tempi di esecuzione di tali operazioni, e si fornisce una valutazione qualitativa dei risultati ottenuti.

5.2.1 Le discipline

Il tempo impiegato nella clusterizzazione del Connected Component è ottenuto prendendo la media dei tempi d'esecuzione di dieci prove di clustering. Questa è misurata in circa 5 minuti su una macchina virtuale in esecuzione su hardware di prestazioni differenti rispetto alle prove precedenti (Intel i7 3.0GHz 8GB RAM DDR3).

Considerando il tempo impiegato per clusterizzare 312'357 nodi, l'operazione sul grafo citazionale risulta efficiente: Chinese Whispers ottiene performance migliori su grafi Small World di grandi dimensioni [Bie06], e la rete citazionale utilizzata nel presente lavoro rientra in questa tipologia.

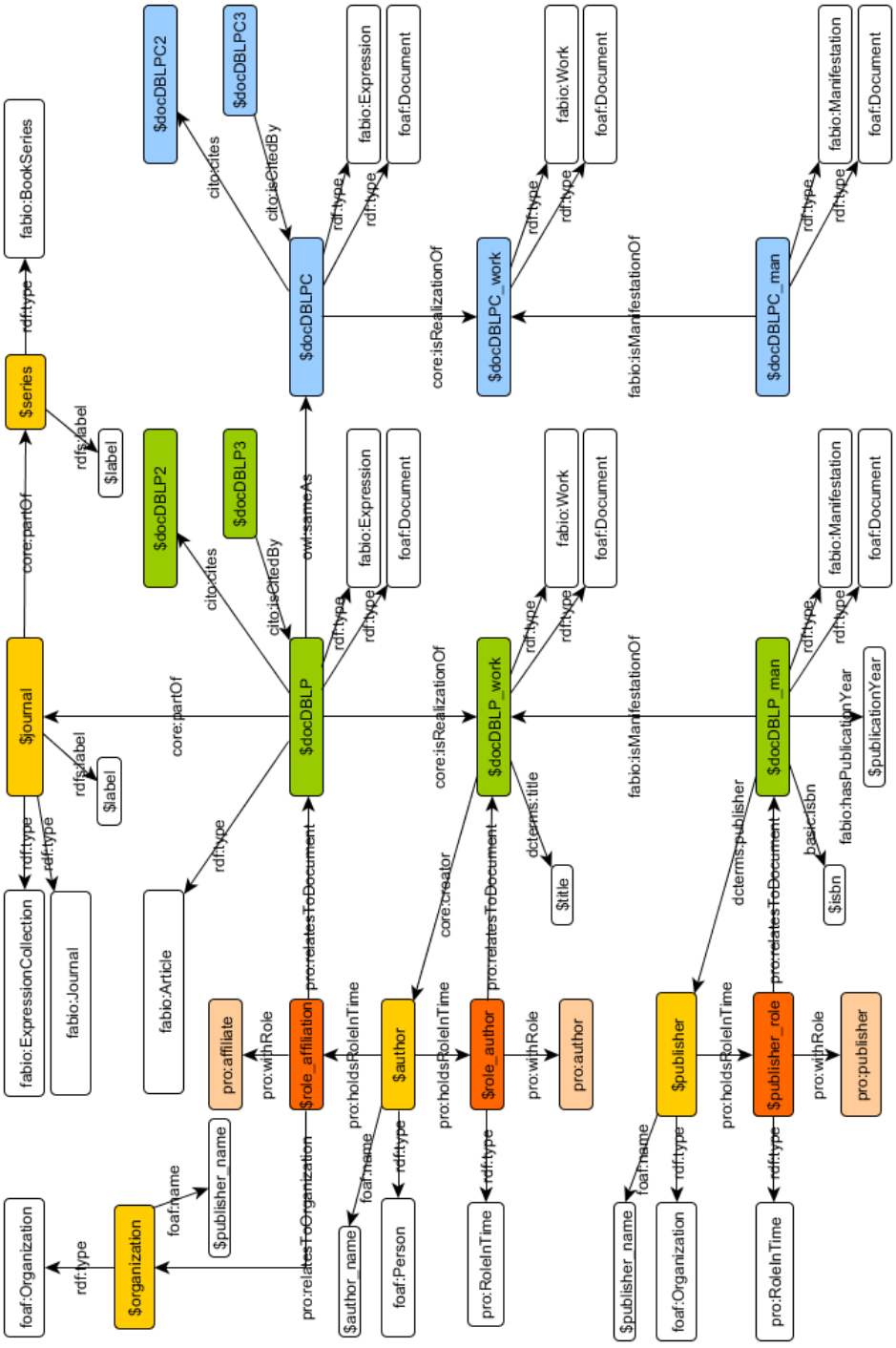

Figura 5.1 La rappresentazione del modello semantico di un articolo nei proceedings di una conferenza.

Dalla tabella 5.1 si osserva tuttavia che la maggioranza dei cluster contiene un numero di pubblicazioni molto piccolo: un effetto che si ottiene quando i nodi (i documenti) al loro interno non riescono a propagare il valore della loro classe al resto della rete citazionale.
Questa condizione si verifica nel momento in cui questi nodi presentano pochi link di basso peso verso gli altri nodi del grafo, e porta ad interpretare i dati a disposizione secondo due valutazioni:

- le pubblicazioni all'interno di questi insiemi non condividono molte citazioni con il resto della rete citazionale. Ad esempio, i nodi di questi cluster potrebbero rappresentare pubblicazioni che citano, ma che non vengono mai citate (da cui consegue la bassa embeddedness di questi nodi rispetto al resto del grafo);

- i record dei dataset utilizzati mancano di alcune voci di bibliografia

cluster contenenti 1'000 o più pubblicazioni	27 ($\sim$ 0,007%)
cluster contenenti 100 $\div$ 999 pubblicazioni	358 ($\sim$ 0,09%)
cluster contenenti 10 $\div$ 99 pubblicazioni	3 819 ($\sim$ 1%)
cluster contenenti 2 $\div$ 9 pubblicazioni	363 639 ($\sim$ 95,6%)
cluster contenenti 1 pubblicazione	12 684 ($\sim$ 3,3%)

Tabella 5.1 Il risultato del clustering del Connected Component.

5.2.2 L'indice di multidisciplinarietà

Come visto nel capitolo 4.4, la multidisciplinarietà di un prodotto della ricerca è calcolata attribuendo "+1" ogni volta che la pubblicazione riceve un numero di citazioni sopra la soglia di rilevanza rispetto alla disciplina citante.

CoData calcola tale soglia analizzando la distribuzione delle citazioni tra ogni coppia di discipline: un'operazione che richiede la creazione di una matrice dei percentili citazionali.

Di seguito si riporta il calcolo del tempo d'esecuzione speso nella creazione della matrice, e si fornisce una valutazione qualitativa dei risultati ottenuti.

Il tempo d'esecuzione necessario per la creazione della matrice dei percentili è dipendente dall'insieme delle discipline considerate: al momento dell'avvio del calcolo della multidisciplinarietà di un prodotto della ricerca, l'utente può scegliere il numero minimo di documenti che un insieme disciplinare deve presentare per poter essere considerato rilevante ai fini del calcolo della multidisciplinarietà.

soglia documentale (numero minimo documenti nell'insieme disciplinare)	tempo medio (secondi)
50	5.099
100	5.452
200	6.367
500	8.975
1000	12.065
2000	18.432
5000	38.516

Tabella 5.2 Tempi di calcolo della matrice dei percentili a diverse soglie del numero minimo dei documenti di una disciplina.

In tabella 5.2 sono riportati i tempi di esecuzione necessari alla creazione della matrice dei percentili, in diverse soglie del numero di documenti minimo per disciplina. I risultati sono ottenuti prendendo la media del tempo di esecuzione di dieci prove per ciascuna soglia.

L'andamento della multidisciplinarietà è monotono non decrescente, e aumenta con l'abbassarsi della soglia del numero di documenti minimo richiesto per ogni insieme disciplinare (ovvero all'aumentare del numero di insiemi disciplinari considerati). Questo è visibile in figura 5.2, dove sono rappresentati i valori della multidisciplinarietà per l'articolo di identificativo "dblpc:dblpcEntry1127831" visto in tabella 4.4.

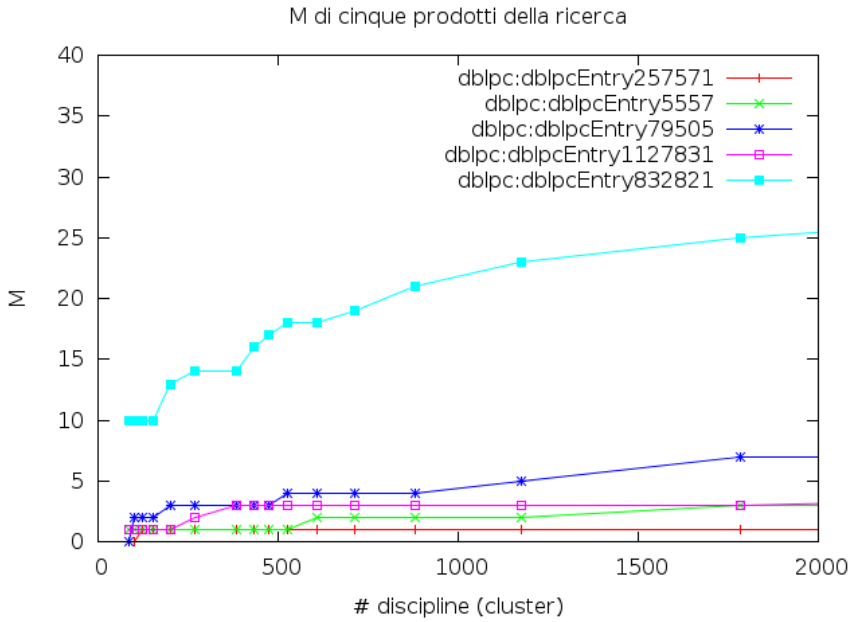

Figura 5.2 Monotonia della multidisciplinarietà di cinque prodotti della ricerca.

L'indice di multidisciplinarietà M risulta complessivamente robusto:

- non è direttamente manipolabile dall'autore di una pubblicazione perché risultante dalle citazioni ricevute;
- non varia sensibilmente all'aumentare delle citazioni ricevute nel contesto di una sola disciplina perché non aumenta all'aumentare dell'impatto ottenuto, ma solo in relazione al superamento della soglia di rilevanza per la disciplina citante;
- dipende dal superamento di una soglia di rilevanza che è calcolata per ogni coppia di discipline, e che segue la distribuzione delle citazioni di entrambe: *nel complesso* delle citazioni che *ogni* pubblicazione della disciplina citata riceve da *ogni* pubblicazione della disciplina citante;
- non decresce nel tempo, e non dipende dall'età del prodotto della ricerca o dall'età della citazione.

Conclusioni

Nel presente lavoro è stato illustrato l'attuale panorama degli indici di valutazione della ricerca, ed introdotto un nuovo indice bibliometrico: l'indice di multidisciplinarietà.

Questo nuovo indice non restringe la valutazione al solo conteggio delle citazioni ricevute da un'opera, ma considera il contesto disciplinare presso il quale la pubblicazione è citata, e valutata rilevante ai fini della contribuzione.

Approcciare la valutazione della ricerca in relazione al contesto disciplinare ha portato al bisogno di creare un modello citazionale più espressivo, capace di formalizzarne le relazioni nel dettaglio, e rappresentare fedelmente il panorama dell'editoria della ricerca.

In questo è stato fondamentale l'utilizzo di alcune tecnologie del Semantic Web, in particolare delle ontologie della famiglia SPAR [PS12; PSV12], le quali si sono rivelate sia in grado di formalizzare le relazioni delle pubblicazioni dell'editoria della ricerca, sia in grado di permettere il riuso delle entità descritte; rendendo più semplici le operazioni di integrazione ed arricchimento di dati provenienti da differenti dataset.
Questo nuovo modello di rappresentazione dei record di un dataset ha per-

messo:

- l'inferenza di nuovi fatti sui dettagli di un prodotto della ricerca;
- una più semplice operazione di integrazione e riuso dei dati e delle loro proprietà.

Tale rimodellazione ha reso possibile un rappresentazione più estesa del mondo della ricerca, senza limitarne i contenuti ad un insieme di journal noti, ma comprendendo ogni forma, aspetto e particolare della contribuzione: le pubblicazioni, gli eventi, i ricercatori.

Da questo modello è stato inoltre possibile osservare la presenza di comunità disciplinari attraverso l'analisi della rete citazionale delle pubblicazioni.

Lo studio è stato condotto sfruttando le proprietà di questa rete, ed abbandonando la classificazione della ricerca secondo le tassonomie delle discipline (troppo generiche, e poco flessibili all'inclusione di nuovi termini) in favore dell'adozione di una tecnica di clustering del grafo delle citazioni.
Questo procedimento si differenzia dai precedenti metodi di classificazione della ricerca in quanto completamente automatico, e non assistito da un processo di gerarchizzazione concettuale degli ambiti della ricerca.

Una definizione di multidisciplinarietà di un prodotto della ricerca è stata quindi formalizzata attraverso lo studio della distribuzione delle citazioni tra la discipline del grafo, e definita entro i termini di una *soglia di rilevanza* citazionale per la disciplina citante. Tale indice di valutazione è risultato complessivamente

robusto in quanto difficilmente manipolabile dall'autore di una pubblicazione, ma dipendente dalla distribuzione delle citazioni nel complesso delle coppie di discipline

equo in quanto non decrescente nel tempo, ed indipendente dall'età della citazione o della pubblicazione.

L'indice così definito non dipende dalle categorie di una tassonomia della ricerca, e non offre una valutazione dell'impatto della stessa, bensí provvede ad una valutazione della rilevanza e della direzione della ricerca di un contributo, proponendo quindi di *integrarsi* all'attuale panorama degli indici bibliometrici, aggiungendo così un nuovo mezzo di valutazione dei prodotti della ricerca.

Il presente lavoro è solo una parte di una più ampia indagine.
Innanzitutto sono state incontrate molte difficoltà nella fase di integrazione di nuovi dati (vista al capitolo 4.2), dovute essenzialmente all'impossibilità di unificare le descrizioni dei record a meno di non ricorrere a complesse procedure di pattern-matching e verifica di corrispondenza dei contenuti.

Questo potrebbe essere evitato con un nuovo approccio all'editoria della ricerca: l'adozione delle ontologie del *Semantic Publishing*, e la pubblicazione dei contenuti nel *Linked Data*, la quale faciliterebbe il processo di riuso, arricchimento, e referenziazione univoca dei dettagli e delle proprietà di un contributo nel *World Wide Web*.

In ultimo, la creazione di un authority file per i titoli ed i nomi delle pubblicazioni, gli eventi, ed i ricercatori, aumenterebbe la precisione dei dati disiponibili, evitando situazioni di ambiguità, o di disparità dell'attribuzione del merito di una contribuzione.
A tal riguardo, un fatto curioso a proposito di quest'ultimo caso è avvenuto proprio durante lo svolgersi del presente lavoro: il titolo di una pubblicazione spesso usata negli esempi (listato 5.1), correttamente rimodellata, ed integrata con le descrizioni di entrambi i dataset DBLP e DBLP+C, presenta un errore di ortografia. Al suo vero titolo,

"Coordinating Multiagent Applications on the WWW: A Reference Architecture."

è infatti sostituita la stringa

"Coordinating Multiagent Aplications on the WWW: A Reference Architecture."

la quale è errata, ma presente in entrambi i dataset.

Questo errore è risultato completamente trasparente al processo di integrazione dei dati, e non è escluso che possa essere anzi occorso anche in altri record, come ad esempio il nome di un autore, o il nome di una publication venue.

Resta tuttavia da valutare se, a causa di questo errore, il conteggio delle citazioni ricevute da questa pubblicazione al di fuori del presente lavoro sia quello effettivo (reale), e non sia stato invece compromesso dalla variazione del titolo del contributo.

Ancora relativamente all'arricchimento delle descrizioni delle pubblicazioni, un futuro sviluppo di questo lavoro dovrebbe considerare l'integrazione di nuovi dataset delle citazioni, come ad esempio *Scopus*[4] e *ACM*[5].

Questo procedimento richiederebbe un nuovo sforzo di pulizia dei dati grezzi, ma verrebbe ricompensato una più ricca rappresentazione del mondo della ricerca, parallelamente alla presenza di un corpo citazionale più completo, il quale permetterebbe anche di comprendere se la presenza di un numero elevato di cluster caratterizzati da un corpo documentale contenuto (attualmente il 98.9% dell'insieme dei documenti), sia conseguenza della mancanza di citazioni, o sia effettivamente una proprietà della rete.

Sul fronte dell'individuazione delle discipline della rete citazionale invece, un futuro sviluppo del presente lavoro dovrebbe ampliare la trattazione degli algoritmi di graph clustering, aumentando il numero di prove, e testando varie opzioni.

Tra questi, una tipologia interessante potrebbe essere quella degli algoritmi di clustering gerarchici, i quali permettono sia un partizionamento della rete citazionale, sia una gerarchiazzazione delle comunità.
Questo approccio ricalca quello delle tassonomie della ricerca, nelle quali gli

[4]http://www.elsevier.com/

[5]http://acm.rkbexplorer.com/sparql/

ambiti ed i termini della ricerca sono organizzati sia orizzontalmente che verticalmente, e potrebbe offrire una valida lettura dei contesti e delle specializzazioni della ricerca.

Parallelamente alla rappresentazione della rete citazionale in discipline della ricerca, si deve anche cercare un metodo che attribuisca loro un nome, e non solo una definizione d'insieme.

Bibliografia

[Ach+13] Elke Achtert et al. 'Interactive Data Mining with 3D-parallel-coordinate-trees'. In: *Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data*. SIGMOD '13. New York, New York, USA: ACM, 2013, pp. 1009–1012. ISBN: 978-1-4503-2037-5. DOI: 10.1145/2463676.2463696. URL: http://doi.acm.org/10.1145/2463676.2463696.

[AH08] Dean Allemang e James Hendler. *Semantic Web for the Working Ontologist: Effective Modeling in RDFS and OWL*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2008. ISBN: 0123735564, 9780123735560.

[Ben01] Stephen J. Bensman. 'Bradford's Law and Fuzzy Sets: Statistical Implications for Library Analyses'. In: *Ifla Journal* 27 (4 2001), pp. 238–246. DOI: 10.1177/034003520102700406.

[Ben+02] Asa Ben-Hur et al. 'Support Vector Clustering'. In: *J. Mach. Learn. Res.* 2 (mar. 2002), pp. 125–137. ISSN: 1532-4435. URL: http://dl.acm.org/citation.cfm?id=944790.944807.

[BHJ09] Mathieu Bastian, Sebastien Heymann e Mathieu Jacomy. *Gephi: An Open Source Software for Exploring and Manipulating Networks*. 2009. URL: http://www.aaai.org/ocs/index.php/ICWSM/09/paper/view/154.

[Bie06] Chris Biemann. 'Chinese Whispers: An Efficient Graph Clustering Algorithm and Its Application to Natural Language Processing Problems'. In: *Proceedings of the First Workshop on Graph Based Methods for Natural Language Processing*. TextGraphs-1. Stroudsburg, PA, USA: Association for Computational Linguistics, 2006, pp. 73–80. URL: http://dl.acm.org/citation.cfm?id=1654758.1654774.

[BMA12] Vicente P. Guerrero Bote e Félix de Moya Anegón. 'A further step forward in measuring journals' scientific prestige: The SJR2 indicator'. In: *CoRR* abs/1201.4639 (2012).

[BOH11] Michael Bostock, Vadim Ogievetsky e Jeffrey Heer. 'D3 Data-Driven Documents'. In: *IEEE Transactions on Visualization and Computer Graphics* 17.12 (dic. 2011), pp. 2301–2309. ISSN: 1077-2626. DOI: 10.1109/TVCG.2011.185. URL: http://dx.doi.org/10.1109/TVCG.2011.185.

[BP98] Sergey Brin e Lawrence Page. 'The Anatomy of a Large-scale Hypertextual Web Search Engine'. In: *Comput. Netw. ISDN Syst.* 30.1-7 (1998), pp. 107–117. ISSN: 0169-7552. DOI: 10.1016/S0169-7552(98)00110-X.

[CN73] Mark P. Carpenter e Francis Narin. 'Clustering of scientific journals'. In: *Journal of the American Society for Information Science* 24 (1973), pp. 425–436.

[Cou97] Neal Coulter. 'ACM'S Computing Classification System Reflects Changing Times'. In: *Commun. ACM* 40.12 (1997), pp. 111–112. ISSN: 0001-0782. DOI: 10.1145/265563.265579.

[DER] DERI. *Documentation for SWPortal Ontology - Semantic Web Portal Ontology*. Ontology developed as part of the Semantic Web Portal Project: http://sw-portal.deri.org/. URL: http://sw-portal.deri.org/ontologies/swportal#Location.

[DH04] Chris Ding e Xiaofeng He. 'K-means Clustering via Principal Component Analysis'. In: *Proceedings of the Twenty-first International Conference on Machine Learning*. ICML '04. Banff, Alberta, Canada: ACM, 2004. ISBN: 1-58113-838-5. DOI: 10.1145/1015330.1015408. URL: http://doi.acm.org/10.1145/1015330.1015408.

[EK10] David Easley e Jon Kleinberg. *Networks, crowds, and markets: reasoning about a highly connected world*. Cambridge University press, 2010. ISBN: 978-0-521-19533-1.

[Fre77] Linton C. Freeman. 'A Set of Measures of Centrality Based on Betweenness'. In: *Sociometry* 40.1 (1977), pp. 35–41.

[FRfBRLACSC98] IFLA Study Group on the Functional Requirements for Bibliographic Records, International Federation of Library Associations e Institutions. Section on Cataloguing. Standing Committee. *Functional Requirements for Bibliographic Records: Final Report*. K. G. Saur Verlag GmbH, 1998. ISBN: 9783598113826. URL: http://www.ifla.org.

[Gar05] Eugene Garfield. 'The Agony and the Ecstasy— The History and Meaning of the Journal Impact Factor'. In: (2005). URL: http://www.garfield.library.upenn.edu/papers/jifchicago2005.pdf.

[Gar06] Eugene Garfield. 'The History and Meaning of the Journal Impact Factor'. In: *JAMA* 295.1 (2006), pp. 90–93. DOI: 10.1001/jama.295.1.90. eprint: http://jama.ama-assn.org/cgi/reprint/295/1/90.pdf. URL: http://jama.ama-assn.org.

[Gar72] Eugene Garfield. 'Citation Analysis As A Tool In Journal Evaluation – Can Be Ranked By Frequency And Impact Of Citations For Science Policy Studies.' In: *SCIENCE* 178.4060 (1972), pp. 471–479. URL: http://www.garfield.library.upenn.edu/essays/V1p527y1962-73.pdf.

[Gil77] Nigel G. Gilbert. 'Referencing as persuasion'. In: *Social Studies of Science* (1977), pp. 113–122.

[GS63] Eugene Garfield e I. H. Sher. 'New Factors in the Evaluation of Scientific Literature, Through Citation Indexing'. In: *American Documentation* 14.8 (1963), pp. 195–201.

[Hir05] Jorge E. Hirsch. 'An index to quantify an individual's scientific research output'. In: *Proceedings of the National Academy of Sciences of the United States of America* 102.46 (2005), pp. 16569–16572. DOI: 10.1073/pnas.0507655102.

[HM11] Stephen C. Hubbard e Marie E. McVeigh. 'Casting a wide net: the Journal Impact Factor numerator'. In: *Learned Publishing* 24.2 (2011), pp. 133–137.

[JPJ14] Sunghae Jun, Sang-Sung Park e Dong-Sik Jang. 'Document clustering method using dimension reduction and support vector clustering to overcome sparseness'. In: *Expert Systems with Applications* 41.7 (2014), pp. 3204

–3212. ISSN: 0957-4174. DOI: http://dx.doi.org/10.1016/j.eswa.2013.11.018.

[JS10] David Jurgens e Keith Stevens. 'The S-Space Package: An Open Source Package for Word Space Models'. In: *Proceedings of the ACL 2010 System Demonstrations*. ACLDemos '10. Stroudsburg, PA, USA: Association for Computational Linguistics, 2010, pp. 30–35. URL: http://dl.acm.org/citation.cfm?id=1858933.1858939.

[Kle00] Jon Kleinberg. 'The Small-world Phenomenon: An Algorithmic Perspective'. In: *Proceedings of the Thirty-second Annual ACM Symposium on Theory of Computing*. STOC '00. New York, NY, USA: ACM, 2000, pp. 163–170. ISBN: 1-58113-184-4. DOI: 10.1145/335305.335325.

[Lab10] Cyril Labbé. 'Ike Antkare one of the great stars in the scientific firmament'. In: *International Society for Scientometrics and Informetrics Newsletter* 6.2 (2010), pp. 48–52.

[LB09] Loet Leydesdorff e Stephen J. Bensman. 'Classification and Powerlaws: The Logarithmic Transformation'. In: *CoRR* abs/0911.3416 (2009).

[LCR12] Loet Leydesdorff, Stephen Carley e Ismael Rafols. 'Global Maps of Science based on the new Web-of-Science Categories'. In: *CoRR* abs/1202.1914 (2012).

[Ley09a] Loet Leydesdorff. 'Betweenness Centrality as an Indicator of the Interdisciplinarity of Scientific Journals'. In: *Journal of the American Society for Information Science and Technology* (2009).

[Ley09b] Michael Ley. 'DBLP: Some Lessons Learned'. In: *Proc. VLDB Endow.* 2.2 (2009), pp. 1493–1500. ISSN: 2150-8097.

[LV06] Loet Leydesdorff e Liwen Vaughan. 'Co-occurrence Matrices and Their Applications in Information Science: Extending ACA to the Web Environment'. In: *J. Am. Soc. Inf. Sci. Technol.* 57.12 (2006), pp. 1616–1628. ISSN: 1532-2882. DOI: 10.1002/asi.v57:12.

[MY06] Lokman I. Meho e Kiduk Yang. 'A new era in citation and bibliometric analyses: web of science, scopus, and google scholar'. In: *corr* (2006).

[Nar76] Francis Narin. *Evaluative Bibliometrics: The Use of Publication and Citation Analysis in the Evaluation of Scientific Activity.* U.S. Dept. of commerce. National technical information service, PB 252 339. Computer Horizons, 1976.

[Per+08] Denilson Alves Pereira et al. 'Using web information for creating publication venue authority files'. In: *JCDL.* 2008, pp. 295–304.

[Pri69] Alan Pritchard. 'Statistical bibliography or bibliometrics'. In: *Journal of Documentation* 25 (1969), p. 348.

[PS12] Silvio Peroni e David Shotton. 'FaBiO and CiTO: Ontologies for describing bibliographic resources and citations'. In: *J. Web Sem.* 17 (2012), pp. 33–43.

[PSV12] Silvio Peroni, David Shotton e Fabio Vitali. 'Scholarly Publishing and Linked Data: Describing Roles, Statuses, Temporal and Contextual Extents'. In: *Proceedings of the 8th International Conference on Semantic Systems.* I-SEMANTICS '12. New York, NY, USA: ACM,

2012, pp. 9–16. ISBN: 978-1-4503-1112-0. DOI: 10.1145/2362499.2362502.

[RM10] Ismael Rafols e Martin Meyer. 'Diversity and network coherence as indicators of Interdisciplinarity: case studies in bionanoscience'. In: *Scientometrics* 82.2 (2010), pp. 263–287.

[RT05] Erhard Rahm e Andreas Thor. 'Citation Analysis of Database Publications'. In: *SIGMOD Rec.* 34.4 (2005), pp. 48–53. ISSN: 0163-5808. DOI: 10.1145/1107499.1107505.

[Sch07] Satu Elisa Schaeffer. 'Survey: Graph Clustering'. In: *Comput. Sci. Rev.* 1.1 (2007), pp. 27–64. ISSN: 1574-0137. DOI: 10.1016/j.cosrev.2007.05.001.

[Sci04] National Academy of Sciences. *Facilitating Interdisciplinary Research*. The National Academies Press, 2004. ISBN: 9780309094351.

[Sho+09] David M. Shotton et al. 'Adventures in Semantic Publishing: Exemplar Semantic Enhancements of a Research Article.' In: *PLoS Computational Biology* 5.4 (2009). DOI: 10.1371/journal.pcbi.1000361.

[SKM06] Antonis Sidiropoulos, Dimitrios Katsaros e Yannis Manolopoulos. 'Generalized h-index for Disclosing Latent Facts in Citation Networks'. In: *CoRR* (2006).

[Sol70] Derek J. de Solla Price. 'Citation measures of hard science, soft science, technology, and nonscience'. In: *Communication among scientists and engineers*. A cura di Carnot E. Nelson e Donald K. Pollock. Heath Lexington Book, 1970, pp. 3–22.

[Sti07] Andrew Stirling. ‘A general framework for analysing diversity in science, technology and society’. In: *Journal of the Royal Society Interface* 4.15 (2007), pp. 707–719.

[Sti98] Andrew Stirling. ‘On the economics and analysis of diversity’. In: *Science Policy Research Unit (SPRU), Electronic Working Papers Series, Paper* 28 (1998).

[Sur+05] York Sure et al. ‘The SWRC Ontology – Semantic Web for Research Communities’. In: *Proceedings of the 12th Portuguese Conference on Progress in Artificial Intelligence*. EPIA’05. Berlin, Heidelberg: Springer-Verlag, 2005, pp. 218–231. ISBN: 3-540-30737-0, 978-3-540-30737-2. DOI: 10.1007/11595014_22.

[Tan+08] Jie Tang et al. ‘ArnetMiner: Extraction and Mining of Academic Social Networks’. In: *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD ’08. Las Vegas, Nevada, USA: ACM, 2008, pp. 990–998. ISBN: 978-1-60558-193-4. DOI: 10.1145/1401890.1402008.

[TZY07] Jie Tang, Duo Zhang e Limin Yao. ‘Social Network Extraction of Academic Researchers’. In: *Proceedings of the 2007 Seventh IEEE International Conference on Data Mining*. ICDM ’07. Washington, DC, USA: IEEE Computer Society, 2007, pp. 292–301. ISBN: 0-7695-3018-4. DOI: 10.1109/ICDM.2007.30.

[Wal04] Thomas Vander Wal. *Folksonomy Coinage and Definition*. 2004. URL: http://vanderwal.net/folksonomy.html.

[WBB10] Jevin D. West, Theodeore C. Bergstrom e Carl T. Bergstrom. ‘The Eigenfactor MetricsTM: A Network Ap-

proach to Assessing Scholarly Journals'. In: *College & Research Libraries* 71.3 (2010), pp. 236–244.

[WS98] Duncan J. Watts e Steven H. Strogatz. 'Collective dynamics of 'small-world' networks'. In: *Nature* 393 (1998), pp. 440–442.

[Zha+10] Lin Zhang et al. 'Subject clustering analysis based on ISI category classification'. In: *Journal of Informetrics* 4.2 (2010), pp. 185–193. ISSN: 1751-1577. DOI: 10.1016/j.joi.2009.11.005.

[Zit05] Michel Zitt. 'Facing Diversity of Science: A Challenge for Bibliometric Indicators'. In: *Measurement: Interdisciplinary Research and Perspective* 3 (1 2005), pp. 38–49. DOI: 10.1207/s15366359mea0301_6.

Ringraziamenti

Questo è stato un grande percorso.
Un lungo cammino i cui dettagli non sarebbe possibile abbracciare e rivelare in breve.

Sarò vago nel ringraziarvi tutti, perché è proprio questa la cosa che voglio fare con poche righe, lasciando a voi i veri significati dietro le mie brevi frasi.

Queste parole non sarebbero mai state possibili senza l'affetto ed il cuore dei miei genitori e dei miei famigliari. A loro quindi va il mio primo ringraziamento, per la pazienza e per il coraggio.

Ai miei amici, vicini e lontani, grazie per le serate ed i tanti caffé. Per la gioia e la passione che mi avete dato. Per la vostra bellezza.

Ai miei professori, ed in particolare al gruppo Bibliometrics, per aver riposto in me la loro fiducia, condividendo l'instancabile curiosità.